万卷方法

实用抽样方法

SHIYONG CHOUYANG FANGFA

■ 加里 · T.亨利 著　　沈崇麟 译



重庆大学出版社
http://www.cqup.com.cn

万卷方法

实用抽样方法

SHIYONG CHOUYANG FANGFA

■ 加里·T.亨利 著　沈崇麟 译

重庆大学出版社

图书在版编目(CIP)数据

实用抽样方法/(美)亨利(Henry, G.T.)著;沈崇麟译.—重庆:重庆大学出版社,2008.8
(万卷方法)
书名原文:Practical Sampling
ISBN 978-7-5624-4487-9

Ⅰ.实… Ⅱ.①亨…②沈… Ⅲ.抽样调查—调查方法
Ⅳ.C811

中国版本图书馆CIP数据核字(2008)第058461号

实用抽样方法
加里·T.亨利 著
沈崇麟 译
责任编辑:雷少波 罗 杉 版式设计:雷少波
责任校对:贾 梅 责任印制:赵 晟
*
重庆大学出版社出版发行
出版人:张鸽盛
社址:重庆市沙坪坝正街174号重庆大学(A区)内
邮编:400030
电话:(023) 65102378 65105781
传真:(023) 65103686 65105565
网址:http://www.cqup.com.cn
邮箱:fxk@cqup.com.cn (市场营销部)
全国新华书店经销
四川省内江市兼升印务有限公司印刷
*
开本:940×1360 1/32 印张:4.75 字数:145千
2008年8月第1版 2008年8月第1次印刷
印数:1—4 000
ISBN 978-7-5624-4487-9 定价:23.00元

作译者简介

加里·T. 亨利(Gary T. Henry) 在美国弗吉尼亚大学公共管理学教授研究方法、统计学、项目评估和政策分析等课程。曾担任弗吉尼亚公立学校教学评估指标体系研究的项目负责人和中国某大学统计学客座教授。此外,他也是国家议会全国会议的成员和演讲人。

早年在威斯康星大学取得博士学位之后,他进入了弗吉尼亚审计和评审委员会,在那里担任首席方法专家,用定量方法进行教育、惩处和交通问题的研究。随后在州内阁担任教育部副部长。他的研究兴趣涉及稳健估计、实地因果关系评估和公正评测等方面。亨利博士是《评估评论》(*Evaluation Review*)和《公共管理评论》(*Public Administration Review*)等杂志的非常活跃的撰稿人和评审人。他也是美国评估协会规范和道德委员会的主席。

沈崇麟 中国社会科学院社会学研究所研究员,中国社会科学院研究生院教授,"第二、三、四批百县市国情调查问卷调查"、"中国城乡居民生活调查"、"中国城乡家庭变迁调查"和"中国城乡社会变迁调查"等研究项目主持人。

致谢

没有诸多同事和同行的指教和帮助，就不会有《实用抽样方法》一书的存在。玛丽·斯图茨曼（Mary Stutzman）、西摩·苏德曼（Seymour Sudman）和格里格·瑞斯特（Greg Rest）的评阅使我受益匪浅。感谢丛书的两位编辑列昂纳多·贝克曼（Leonard Bickman）和黛博拉·罗格（Debra Rog）的鼓励和盛情邀请。肯特·迪克西（Kent Dickey）和弗吉尼亚·黑廷格（Virginia Hettinger）认真校阅了文字、公式和计算结果。堤姆·亨德里克（Tim Hendrick）和帕特·斯特里诺（Pat Storino）费心为各种统计图润色。休·马汉（Sue Mahan）和安妮·科尔茨（Annie Kurts）做了大量的文字整理工作。最后，我还要感谢妻女的信任和支持。她们从不抱怨我将周末用来写作，而不是修理家中的庭院，她们也从来没有怀疑过，我的付出终将得到回报。

目录

第1章 导论

Introduction

大多数用于社会和政策科学的数据都收集自样本。公众舆论调查、社会实验和教育改革评估等方面的研究一般都要采用抽样方法。任何须将从调查的对象或单位得到的结论外推到更大的研究总体的规范的调查研究都要使用抽样方法。我们经常对样本进行调查，很少对整个研究总体进行调查。诸如十年一度的人口普查这样的总体调查是不多见的。收集用作项目评估的数据若不以抽样方法为依据，那么新的教学方法或社会服务设施的交付使用的风险和费用是很难估算的。同样，如果没有抽样方法，我们也无法对新项目的效力进行科学的评估。

在20世纪末叶，概率抽样技术的引进和使用极大地推动了经验的社会和政策科学研究的发展。尽管抽样方法是如此重要，但是社会和政策科学领域中的研究者却很少有这方面的实际经验，因而十分希望能得到有关抽样的实际应用的指导。有关抽样方法的一些假设常常构成了研究者希望采用的分析技术的基础。研究者经常都需要证明那些假定都是成立的。抽样常令我们处于两难的境地。一方面，限于时间和经费，研究者无法收集一个特定的研究所感兴趣的整个群体或总体的数据。然而另一方面，研究者和研究成果的使用者通常所感兴趣的是总体，而非总体的某一子集。正因为如此，克服时间和费用的限制与我们对于总体的信息的需求这一矛盾，即将来自总体的某一子集或"样本"的研究发现推广到整个总体便变得如此的关键和重要。而有关抽样的逻辑和方法的知识正是确保我们是否能将来自研究的参与者或被调查人的结论，合理地推广到他们所代表的总体的基础。

对抽样方法和它的基本含义有所了解，这一问题在政策研究中尤其重要。在政策研究的实施和使用过程中，所遇到的问题通常都与抽样设计方案的选择问题有关：政策和项目所定义的目标总体是否与研究所定义的相同？我们使用的选取对象和单位的方法是否会使决策者所需要的估计值产生偏倚？来自样本的估计值，就研究的目的而言是否已经足够精确？表 1.1 简要地列出了这三个问题，用于判断研究的有效性的标准和研究结果使用的潜在含义。

我们将用一个有关高龄老人的服务需求的评估的例子，来揭示不同的抽样设计对评估结果的影响：

- **总体的定义**。本研究旨在对某州全体高龄老人的服务需求进行评估。如果样本抽取自一个由目前正在接受各种公开提供社会和医疗服务的机构的服务的高龄老人组成的总体，那么那些目前没有接受这些机构提供的服务，但需要这些服务的老人将被排除在样本之外。这样的选择势必会导致对实际需求的低估。
- **抽样方法**。采用一种只关注居住在集体宿舍（在这里集体宿舍指养老院、疗养院等，而不是我们通常理解的那种集体宿舍，译者注）中的老人的抽样策略将会使结果有偏倚。在许多情况下集体宿舍提供的照料水平都是最高的。因而在将这样的需求的估计值外推到州的整个的老龄人口时，可能是偏高的。这一样本可能没有包括足够的独自居住的或与自己的家人生活在一起的自己照料自己的老人。
- **估计的精度**。来自样本的估计值，例如均值或比例数，即使在抽样方法是无偏的时候，也是会有一定的波动的。一个未被告知实情的研究结果的用户可能认为估计值是严格而精确的，因而对“精确”的估计值过于信任。例如，一个来自需求评估的估计值显示，有 63.4% 的高龄老人需要服务。而对于一个小样本来讲，研究者可能有理由认为真的均值在 54% 到 73% 之间，假定在一个 500 000 名高龄老人的总体中，服务人口数的估计值高低的波动范围，几乎达到了 100 000。这说明估计值是很不精确的，因而我们也无法据此来制定相应的政策和建立为高龄老人提供服务的项目。

表1.1 政策研究中的样本设计问题

问题	判断标准	含义
总体的定义	目标总体和研究总体的一致性	因为将那些并非目标总体的成员包含在内,或遗漏了目标总体的成员而导致的研究总体的偏倚
抽样方法	选择样本时,给研究总体中的任何成员以相等的选择的可能	如果研究总体中的某些成员的被选择的可能性比其他成员更高,抽样方法便将导致结果的偏倚
估计的精度	估计精确度足以满足政策制定的需要	所有的样本产生的估计数都不是确定的数字。缺乏精确度势必影响政策的制定

定义样本

在科学和日常语言中,"样本"一词常以各种不同的方式使用。例如马克和沃克曼(Mark and Workman,1987, p. 47)指出:"对化学家来讲,样本一词也许会令他们联想起一堆东西,一个盛有某种液体、糊状物的容器,或一堆难以言状,但成分却是确定的东西。"对于化学家来讲,样本具有内在的重要性。化学家手头得到的那些东西,例如一个犯罪现场或从某种较大的东西上选取的一部分可能就是全部问题之所在。对化学家而言,他们的根本目的就是确定样本的成分。化学家在法庭上的职责是发现茶中的砒霜,而不是茶的总体的代表。

化学家的样本可被视作一种标本,因而特例是很重要。相反,在研究文献和本书中,样本则是指总体中用来得到整个总体的信息的一个子集。在这个意义上讲,样本是总体的模型。一个上佳的样本将会完好地代表总体。社会或政策科学家的真正兴趣并不是样本——样本只不过是他们用来了解总体的工具而已。

以上的讨论自然而然地会产生以下两个问题:

- 如何选择一个用于代表总体的样本？
- 如何判断样本是否完好地代表了总体？

有关第一个问题的基本要点我们将在下一章进行讨论，此后，这一讨论将贯穿全书。在第3章，我们将对样本和样本设计过程中选择的抽样框中可能存在的误差源进行扼要的介绍。

在我们转入这些问题的讨论之前，有必要对我们使用的“有代表性”一词做一些解释。一个样本是用来代表总体的，因此它是总体的一个模型或代表。我们将“有代表性”这一术语加在样本上，就如我们日常所言的“有代表性的样本”一样，并未给样本提供任何额外的信息。形容词“有代表性”并不是技术上的定义，只不过是这一术语的使用者的一种主观判断而已。我们并未因此而建立确定一个样本究竟是有代表性的还是没有代表性的标准。尽管如此，“有代表性的样本”仍然是人们给自己所提供的样本所做的最为常见的描述。为了避免不必要的误解，我们建议在样本和总体之间的对应问题的描述上去掉这一形容词，改为采用一些能对样本选择过程进行描述的词。有关这样的描述的重要性，我们将在随后的一章进行讨论。

抽样和效度

很少有研究者能在一个特定的研究中，收集所有自己感兴趣的对象的数据。而样本则为我们提供了切实可行和有效地收集数据的工具。样本好比总体的模型，然而研究者若要将研究的发现推广到总体，则必须要使模型是总体的精确的代表。

研究者或某一项研究成果的使用者将研究的发现推广到样本个体、时间和地方之外的能力称作“外部效度”（external validity）（Campbell & Stanley，1963；Cook & Campbell，1979）。库克和坎贝尔认为外部效度的关键问题是“如果从建构A到建构B可能存在某种因果关系，那么在跨越不同的人、场合和时间时，这种关系究竟在多大程度上可以被普遍化呢？”（Cook & Campbell，1979，p. 39）例如，一些研究者发现，在三年级阅读课程中使用计算机辅助教学软件包，能使位于市中心的学校的学生的词汇量和理解力有所提高。这一研

究成果的用户有理由提出一连串的问题:诸如这样的教学方法是否在农村地区的学校也同样有效?在四年级阅读课中是否也有效?学习成绩的提高是否源自计算机使用的新奇?一旦学生不再对计算机感到新奇,学习成绩便不再会因为它们的使用而提高?

对研究的发现进行概括的能力是数据实际从中收集的样本的一个功能。抽样设计和实施都会对诸如这样的概括能力有所影响。本书所介绍的实用抽样方法对设计和实施二者予以同样的重视,因为这二者都可能对研究的效度或总误差有所影响。样本设计的内容包括:挑选恰当的选样方法,如随机数码拨号法;确定研究所需的个案数目等。将一个设计付诸实施则包括:得到有关研究总体的完整的清单;收集真实可靠的数据;确保调查答案的确收集自一个其组成能精确地代表总体的群体。任何有可能影响数据实际收集的群体的组成的计划或行为,都关系结果的概括与推广。正因为如此,切实可行的抽样设计必须与整个研究设计和实施整合为一体。

除了外部有效性之外,我们还必须考虑其他两种统计效度问题。这两种统计效度会直接对样本设计有所影响,反过来,样本设计也会对它们有所影响。这两种统计效度问题的讨论见诸于库克和坎贝尔的有关著作(Cook & Campbell, 1979)。统计结论的效度是一种得出有关出现在样本数据中的关系,即所谓的共变的结论的能力。通常我们用统计检验来考察观察到的关系是否由变化引起的,或者如克雷默和西曼所言:"在将统计检验运用于数据时,它定义了一条规则,一条是否能拒绝零假设(null hypothesis),即证据是否是不容置疑的规则。"(Kraemer & Thiemann, 1987, p. 23)因为这些检验对关系的大小(效应的大小)和样本的大小二者都十分敏感,所以样本的大小可能对避免得出"伪的共变结论"是至关重要的(Cook & Campbell, 1979, p. 37)。

在进行统计检验时,小的样本量可能导致某种保守的偏倚(第二种类型的误差)。当实际上零假设是不成立的,但却未能被否定的时候,第二种类型的误差便会发生。在遇有这样的情形时,一个被检验的项目或干预政策将被认为是无效的,尽管实际上它是有效的。然而"合理的怀疑"标准可能无法满足给定的期望效应量和实际的样本量。在发生保守的偏倚时,较小的效应或共变虽然的确存在,但是样本量却不足把效应记录到统计显著性的界限之内。在对一个参与人

数少和样本量比较小的试行项目进行评估时，发生这样的现象尤其令人感到沮丧。项目产生的虽然不太大，但却是很有意义的效应，可能会因为缺乏统计的显著性而被吞噬。这样，评估者可能就会因此而错误地认为项目是无效的。我们将会在第7章对样本容量和统计结论的效度问题进行讨论（希望对这一问题有更多了解的读者，可参见 Kraemer & Thiernann，1987；Oakes，1986；or Lipsey，1989）。

有关统计结论的效度的第二个方面的问题——量度的信度（reliability of measures）是又一个我们在抽样设计和实施时应该考虑的问题。数据中观察到的变差越大，我们所使用的工具的信度就越低（Cook & Campbell，1979）。每当观察到的变差加大的时候，尽管真关系的确存在，虚无假设的否定也会因此变得更加困难。假如我们使用的工具是无偏的，那么比较大的样本量可在一定程度上对加大的变差有所补偿。然而为了补偿因为工具的信度的缺乏而导致的方差的波动，我们必须在设计阶段的初期就对这一问题有所了解。

为什么要抽样

抽样涉及的信度问题既然如此复杂，那么研究者就可能会因此而提出“我们为什么要抽样？”这样一个问题。其实，抽样无非是达到我们最终目的的一个切实可行的手段。研究者的工作通常都始于一个目标总体。而这一总体一般都取决于一个政策或项目。研究者通常都会对有关政策或项目提出一个问题。例如，一个研究者可能会提出“为处于风险中的四岁儿童（at-risk 4-year-olds）建立一个学前教育项目是否会提高这些儿童的认知能力，从而减少他们在以后的学习中对特殊教育援助的需求？”这样一个问题。如这样一个由制定政策的人确定的处于风险中的四岁儿童的目标总体将被纳入我们研究的问题中。

通过抽样，研究者将这一问题转变成一个可行的经验项目。不言而喻，我们不可能为所有处在风险中的四岁儿童提供这样的发展项目，并在他们身上进行数年的测试，以确定项目究竟有什么影响。限于资源，我们无法这样做。在这一例子中，我们受到了来自两个方面的条件的限制——项目方面和研究方面。一方面，寻找项目发展

所需的资助、设备和训练有素的工作人员就是很困难的。另一方面，得到数据收集与分析和随后进行的评估等需要的投资同样也不是一件容易的事。而耗费很多的公共资助，却未能对项目的影响做出评估的做法显然是很不妥当的。

抽样使我们得以使用总体的一个子集来对一个项目进行测试。我们之所以要进行抽样的主要原因固然是由于研究项目的资源是有限的，但是，同时也因为抽样也能使研究的质量有所改进。例如，由于能完全胜任三、四岁儿童的预调查的训练有素的研究人员人数有限，我们不得不聘用训练不足的工作人员，或对风险状态进行过于简单的测试，因而无法得到可靠的结果。而抽样则可以使我们得以集中使用资源，从而提高每一个工作人员收集的数据的数量和质量，同时最大限度地减少数据的丢失。

在有些场合，研究人员可能会发现抽样方法是不可取的。这种场合可能有很多，我们无法一一列举，但其中两种情况我们希望读者能够记住：小总体和有可能降低结果的可信性的抽样。在处理小总体时（成员数低于50），从整个总体收集数据会提高数据的可信度和可靠度。小样本数据中的单个极端个案或离群数据（outlier）的影响尤其显著，因而用总体数据来检验假设会使问题变得更加简单。同样，如果研究项目的用户了解到一个“独特的个案”已从样本中删除，那么结果的可信性同样会受到影响。诸如这样的问题大多发生在小总体中。因为这时研究项目的用户对总体单个成员的信息有着比较详细的了解。

在一个有关如何合理分配公共基金的研究中，如果抽样方法不当，有可能对研究结果的可信性产生不良的影响。例如，使用一个行政区域的样本——城市和郡县的样本来检验地区性特征之间的关系和对基金的需求，虽然可能有助于提高统计的效率，但在政治上却是不可取的。不仅如此，如果在分析中没有体现研究项目的用户（在本例中是立法者）所在的地区，那么研究结果的可信性在他们的心目中就会大打折扣。因为研究者无法明确地阐述，不会因立法者的家乡未在研究中得到体现而导致研究结果的可信性下降。在本例中，由于某一地区的缺失而使某些人趁机对项目提出质疑，导致研究结果可能因此而被弃之不用。

在为一个其结果将在政治环境中使用的研究项目进行抽样时，

我们还应在考虑科学的信度问题之外,做深一层的考虑,即,考虑一下有关政治的信度问题。我们这样说,并不意味着政策研究是不需要抽样的。

在不进行抽样的场合,仍然会涉及一些与抽样设计有关的问题。例如,要得到一张目标总体的全面的清单便是抽样涉及的首要问题。对目标总体的所有成员进行调查的普查同样也需要一张全面的清单。研究中与目标总体的覆盖程度和无回答(即数据丢失的个案)有关的问题,与使用抽样方法和不使用抽样方法这两种研究都有关。

本书概要

本书使用的“实用抽样设计”一词,其含义已经超出抽样理论的范围。实用抽样设计包括抽样理论、设计的逻辑和设计方案的实施。样本设计的逻辑和它的实施贯穿整个研究的所有方面和全过程。研究的性质、测量指标和测量工具、数据收集方法、研究总体的定义和数据收集的方法等都会受到抽样方法的影响。实用抽样设计必须与整个研究方法整合为一体,以便尽可能地提高研究结果的效度。

重要的问题在于,对实用抽样设计问题的思考,必须联系目前人们对抽样设计问题的具体认识。我们将用在抽样顾问和研究团队共事之初,比较普遍发生的一些问题来对这一问题加以形象的阐述。通常研究团队一方开始都会问:“为了将样本推广到总体,我们需要一个多大的样本?”而抽样顾问则会反问:“你们研究的问题是什么?研究的对象是谁?”研究团队就抽样所提出的问题很多时候仅限于样本的大小。但为了提高效度和降低总误差,我们有关抽样问题的考虑决不能仅限于这一点,而必须设计研究的所有方面,贯穿整个研究过程。

本书所介绍的样本设计的思路,为我们提供了在整个研究过程中如何在若干备择设计方案中进行抉择的基础。我们将这一思路称之为实用抽样设计方法。我们之所以使用“实用”一词是因为我们为读者介绍的框架,强调的是各种可供我们选择的方案和如何在这些备择方案中进行选择的操作要领而非抽样的理论。本书将在概念上对这样的框架进行阐述的同时,还辅之以来自实际的抽样实践的详

尽实例。虽然本书并不打算过多地从理论和数学的角度来介绍抽样问题,但它介绍的内容都以以往的理论和数学的抽样著作为基础,而这些著作将为那些有志于对这方面文献作更深入了解的读者提供了一个完整的参考书目。

本书的主要阅读对象是那些将抽样作为自己的研究工具的研究人员。同样,本书也可作为社会和政策科学的本科生的方法课的补充教材,以帮助那些有志于从事研究工作的学生学习有关抽样的知识。除上述两个用途之外,本书也可用作那些在计划自己的研究项目时,在抽样方面需要一些咨询的研究人员的参考书。而那些计划进行抽取大的、复杂的样本的研究人员,则最好去寻求有经验的抽样专家的帮助。

本书第2章主要介绍了两种选择样本的方法——非概率抽样和概率抽样。我们通过若干种基本的设计方案对每种方法进行了介绍。第3章介绍实用抽样方法。它包括两个相对独立的部分:一部分讨论概率抽样中的总误差,另一部分则概要介绍了实用抽样方法的基本路数。这两个部分合在一起则会使我们进一步了解,为什么抽样设计必须要贯穿于一个研究的设计和实施的全过程。实际可供我们选择的方案,在这些方案中进行抉择,以及考虑到其他的抉择的可能性都是实用抽样设计方法的不可或缺的组成部分。

第4章详细介绍了摘自研究文献中的四个例子。这些实例是在第3章介绍的实用抽样方法的框架下组织起来的。这些例子介绍的各种研究,在总体类型、数据收集方法和样本设计上显示出了很大的差异。

其余三章则通过具体例子对实用抽样方法进行了进一步讲解。第5章主要介绍抽样框的问题,第6章则主要介绍抽样技术,而第7章讨论的问题是样本的容量。最后一章则对有关抽样后选择(postsampling choices)问题进行了讨论。

第2章 样本选择方法

Sample Selection Approaches

样本选择方法可归结为两类：概率抽样和非概率抽样。概率样本是以这样的方式，即总体的每一成员实际上都有一个被选入样本的概率选取的。非概率样本的选取都基于研究者的主观判断。研究者一般都以手头进行的研究所要达到的特定目标作为判断的依据。本书主要介绍概率抽样。其原因在于概率样本使我们能对它们进行严格的统计分析，从而确定可能的偏倚和误差。非概率样本则不具备类似的优点。但是在某些场合，非概率样本也不失为一种颇为有用的工具。

非概率抽样

很多研究项目都使用非概率样本。这些样本可以方便地或以某种系统的标准为根据选取。非概率样本实际上是一组性质各异，用主观判断选择的样本的抽样方法的总称。这些方法在确定将总体中的哪些单位选入样本时，都以主观判断为准。非概率样本的选择方法与概率样本的不同，后者都采用基于某种随机机制的选择方法，以确保样本的选择能独立于主观判断。

常用的非概率样本设计有以下六种：

- 方便样本（Conveniences samples）
- 最相似/最不相似样本（Most similar/most dissimilar samples）
- 典型个案样本（Typical case samples）

- 关键个案样本(Critical case samples)
- 滚雪球样本(Snowball samples)
- 配额样本(Quota samples)

表 2.1 概括了这六种非概率样本的性质。

表 2.1 非概率样本设计

抽样类型	选样策略
方便	个案的选取主要考虑它们是否便利于研究
最相似/最不相似	选取条件相似或条件很不相似的个案
典型个案	选择那些事先已经了解有用处,且不极端的个案
关键个案	选择对整体的肯定或评定有举足轻重的个案
滚雪球	由一组成员来确定其他选入样本的成员
配额	调查员选择的样本在某些比较容易认定的变量上有与总体相同的比例。

方便抽样。方便样本是由一组乐于接受调查的个人组成的样本。例如,对电影中的暴力和美国民众的攻击性行为这两者之间的关系感兴趣的心理学家,可以在实验中使用一些来自选修心理学导论课的学生的志愿者作为实验对象。学生首先自愿报告他们自己对暴力问题的态度、倾向,再对他们在一个人为编排的冲突场景中的作为进行观察。随后我们用随机分配法(random assignment)将学生分成两个组。随机分配法是一种用来将自愿参加实验的学生分为一个实验组(treatment group)和控制组(control group)的专门技术。它不同于随机抽样。随机抽样是一种关乎整个样本,即两个组的所有成员的选择的概率方法。我们在本例中使用的方便样本不是随机抽样。在事后将样本分为两个组的时候,譬如像本例那样,我们应该使用随机分配法。

一组成员被示以有暴力画面的电影,另一组放映的电影则没有暴力画面。随后我们对两组的成员进行访谈,再观察他们在可能有暴力冲突的场景中的言行。然后对两组成员在电影放映前后的态度和行为的差异进行比较。

参与这一实验的学生便是一个方便样本。他们是一个现存的群

体的一部分,研究者很方便从这一群体中获取志愿者。获取数据的方便性只是研究者诸多目的中的一个目的。研究的主要目的还在于了解暴力电影对美国民众的影响。但从这样的样本得到的数据究竟能使我们对这一问题有多少了解,则另当别论。

除了暴力电影揭示的因素之外,在理论上讲,还可能存在其他滋长暴力倾向的因素。这些其他的因素是一些混淆变量(confounding variable)。这些变量使我们难以确定自变量(independent variable)(本例中的暴力电影)的影响。例如,个人在精神或情感上的压抑程度可能被假设与暴力倾向有关。不仅如此,年轻人性格可能比较外向,因而更易具有暴力倾向。

在这个意义上讲,样本中的所有学生(包括实验组和控制组)较美国的成年人口总体的压抑程度不是更高,就是更低。就存在于实验组和控制组的攻击性行为的差异而言,压抑程度同样也可能是一个混淆变量。如果学生的压抑程度低于成年人口总体,那么我们实验得出的暴力电影不会使暴力倾向有所增长的结论,并不适用于成年人口总体。这样的设计的主要问题在于在问题涉及压抑程度时,使用这样的样本作为总体的模型是不够精确的。不仅如此,这样的样本也无法了解其他假设的可能引发暴力行为的因素。因此,我们无法断定暴力电影的影响(或没有影响)这样的结论是否适用于样本之外的场合。

在本例中,年龄是一个更易观察到的混淆因素。大专院校的学生年龄大多数在18到25岁之间。如果年轻的成年人更倾向于暴力行为的话,那么样本就可能有偏向更高的侵略性行为的偏倚。选择如本例所示的方便样本,可能会使结果有偏倚,因而使我们无法将得出的结论推广到总体。不言而喻,其他因素也可能对暴力态度和行为有所影响,而这些因素在样本和总体之间的分布是不均衡的。

不确定性和偏倚可以有两种方法加以控制:一种方法是使用更多的数据,对方便样本加以改进;另一种方法是放弃使用方便样本,改而使用概率样本。首先,对那些易于量度的特征,如年龄,我们可以直接加以控制。一个可供我们选择的方便样本是比例代表样本,即按每一年龄组的比例(但可能不是从一个心理学班)来分配和选择样本成员。

我们难以得到有关总体的压抑程度方面的信息。而这些信息对

于分组和按比例地选择个体都是必不可少的。随机选择可以确保总体的每一成员都有被选入样本的可能,给我们提供了一个控制压抑程度的量度。随机选择是一种隐含的控制,用来得到多种特征,包括压抑程度和其他尚未直接控制的特征的混合体。随机选择,作为概率抽样的一个重要性质我们将在以下节进行介绍。

我们有必要花费一点时间给大家介绍一下本例中将学生随机地分配给实验组和控制组的问题,以使大家对与内部效度(internal validity)有关的选择偏倚有所了解(Cook & Campbell, 1979)。这就是说,实验组和控制组(二者均由来自该校心理学选修班的学生组成)之间的在侵略性行为上存在的差异可归结为暴力电影的影响。随机分配的目的在于去除那些任何其他也许可以解释存在于志愿者身上的差异的变量。然而随机分配并不能解决在将从样本数据发现的效应外推到美国公众时,可能发生的不确定性和偏倚的问题。在本例中,由于方便样本的使用,使我们对样本的发现的概括受到了很大限制。

有关无家可归者人口和非机构(deinstitutionalization,指 20 世纪末发生在美国社会中将精神病患者从公立医院转到社区医疗服务中心的趋势,译者注)问题的文献为我们提供了一个采用方便样本的具体例子。随着政策制定者对无家可归者问题的严重性认识的深入,他们也越来越急需我们能为他们提供有关的经验数据。这就意味着更多的数据将使用方便样本收集(Burnam & Koegel, 1988)。通常教堂和其他各种为无家可归者提供床位的场所都是我们用来选择调查对象和收集数据的地方。因为有关什么样的人才可以得到床位的规定将某些人排除在了样本之外,所以这样的样本可能存在某些偏倚。此外,某些无家可归者并不愿意,或不能通过得到一张床位所要求的手续。

近年来,有两个有关无家可归者的研究,试图为我们提供一些基于概率样本的信息。罗西、赖特、费希尔和威利斯(Rossi, Wright, Fisher, and Willis, 1987),以及伯纳姆和考杰(Burnam & Koegel, 1988)描述的方法,虽然比方便样本需要更多的时间和预算,但是他们的发现确实与以前的调查有所不同。例如,伯纳姆和考杰发现只有不到一半(44.2%)的无家可归者在教堂或收容所的床上过夜,而基于方便样本的这一数字几乎为三分之二(66%)。二者存在明显的

差异。

最相似/最不相似个案抽样。最相似/最不相似样本由立意样本(purposeful sample)变化而来,多用于比较性的政体研究和政策取向的个案研究。为了比较政治、社会和经济体系之间的关系,西方国家,如美国、加拿大、英国、法国、德国和意大利经常被组合在一起作为一个国家的样本。其他的组合,例如发展中国家组合,则常被用来进行诸如国家债务的增加对国家的生活水平的影响这样的研究。在最不相似设计中,为了对政策的实施结果进行比较,个案研究经常选择"最好"和"最坏"的个案。

如果个案和资源都比较有限,而所需的信息主要用来进行比较,这些研究是很有用的,但是问题在于,由这样的研究得到的结论能否推广到个案研究的范围之外。在考虑财产税率或断定州资助是否适当时,一个地区常常需要将自己与相邻的地区进行比较。在这种情况中,地理上的邻近程度是相似性的操作定义。这些比较既可能,也可能不太精确地刻画那些在课税基数或提高地方收入能力相似的地区的相对位置。

典型个案抽样。在时间和资源极度不足时,我们也常常会改而选取非概率的样本。"典型个案"设计便是在遇有这样的情况时采用的抽样设计。在采用这样的抽样设计时,研究者选取为数不多的几个、他们认为是正常的或一般的个案。为了提高设计的信度,那些看做独特的或特殊的个案都不会被选入样本。

在这样的设计中,研究者本人对总体的判断和了解对于样本的信度是至关重要的。在政策研究中,"典型个案"样本的选取常常需要进行非常严密的详细审查。对选择偏倚的疑虑是这样的样本存在的一个普遍的问题。使用"典型个案"会引起人们对选取的个案进行详细的审视,而在人们感到个案并非那么典型时,我们发现的那些信度就会因此而大为减色。这样一种类型的设计的个案选择法,其注意力主要集中在每一单独的个案上。样本变成了审视的焦点。

关键个案抽样。另一种在许多方面都与"典型个案"设计颇为相似的非概率样本设计是"关键个案"设计。在关键个案设计中,研究

者选择了数目有限的几个个案,这些个案在逻辑上或依据以往的经验允许我们进行总体的推论。为了预测大选的结果,我们可以关键个案的逻辑为依据来选择关键的选区。在美国社会几乎是家喻户晓的格言"缅因在握,美国在握"(As Maine goes, so goes the nation),便是这一设计应用的极端例证。在 1948 年前的美国大选中,用缅因州的选举结果来预测全美的选举结果是非常灵验的。1948 年的大选中,缅因州的选民大多数都把票投给了杜威(Dewey)。于是有人喊出了这样的口号"见缅因而知佛蒙特"(As Maine goes, so goes Vermont)。

滚雪球抽样。与其他非概率样本显著不同的非概率样本是滚雪球样本。滚雪球抽样借助先前确定的一组成员来确定总体中的其他成员。随着新近确定的成员不断列举其他的成员,样本如同滚雪球一般逐渐成长壮大。滚雪球抽样多用于不存在总体的清单,且研究者也无法自行编撰诸如这样的清单时。有关各种不同的群体,如非法侨民和社区"权力精英"的社会学研究常采用这种方法来生成总体的样本。

配额抽样。另一种非概率抽样得到的样本是配额样本。配额样本将我们研究的总体的群体分为子群体,例如分为男性和女性,或白人、西班牙裔、美国印第安人和其他少数民族等。然后,根据在最终的样本中所需的子群体的比例,分配给调查员既定的、需要他们选取和调查的每一子群体的样本单元数。配额样本与概率样本,特别是与分层样本有许多相似之处,但是它们在一个重要的方面却不尽相同。配额样本允许调查员在选取样本个体时做主观判断。我们给调查员以明确的指示,告诉他们我们希望应选作调查对象的特征。例如,我们可能告诉调查员,在某一邻里地区选取的被调查人数,其中白人和黑人各多少,或在分配好的额度中,男性和女性各多少。在通常情况下,确定的各个子群体的数目将使整个样本中各子群体所占比例与总体相同。

但是我们必须要指出的是,配额样本的被调查人是由调查员选择的。斯图加特(Stuart, 1984)指出,配额样本可能会引起三种问题:

> 正如我们将要从这样一种调查员的自由所证明的那样，在群体抽样中，总是存在着发生选择的偏倚的危险，因为在这样一种抽样中，选择程序的定义并不是十分明确的。至今我们仍然没有可以用于估计这样的样本的标准误差的有效方法……**配额抽样隐瞒了这样那样的无回答问题。**（黑体系原作第43页所强调的要点）

我们之所以说配额抽样隐瞒了无回答问题，是因为每当调查员在遇有拒绝接受调查的被调查人，或找不到户中的任何人的时候，只是简单地选另外一户进行调查而已。因而调查员总是可以得到要求的调查数，但是总体中那些难于联系到的被调查人的比例却可能会因此而被低估。

库克和坎贝尔建议采用一种特殊的配额抽样法。他们把这种方法叫做"异质详析抽样模型"（the model of deliberate sampling for heterogeneity）（Cook & Campbell, 1979）。他们提出的这种配额抽样策略，要求我们从各种各样可能对调查结果有影响的背景和条件中来选取样本成员。"因此一个综合性的教育实验设计，必须包括来自城市、小镇和农村地区的不同家庭背景的，在天资和价值观上存在很大差异的男孩和女孩。"（Cook & Campbell, 1979, p. 75）而这样一种设计的不足之处在于它难以用来进行归纳概括。研究者至多只可以说"至少从一个样本来看，影响是存在的（或不存在的）"（Cook & Campbell, 1979, p. 76）。虽然在这样一个层次上，我们的陈述可被认为是正确的，但是用一个不同的样本，某些混淆变量（confounding variables）有可能使我们的发现不复存在。不了解样本的前提条件，配额抽样对于理论构建的用处是很有限的。此外，将发现限于某一个特定的样本的做法，会使我们提供的信息对于政策所产生的影响变得十分有限。

鉴于上述理由，配额样本作为一种抽样方法，并不为我们看好。尽管如此，我们一直还在使用配额抽样，因为在通常情况下，它的费用比概率样本低。在以户为基本研究单位的研究中，我们有时会使用一种与配额样本多少有些相似，但实际上却是一种概率样本的抽样方法。带有配额的概率样本要求调查员在特定的地理位置，按照特定的路线，对配额规定的数目的被调查人进行调查。

苏德曼对我们之所以将这样一种抽样形式看做一种概率抽样的根据做了简要的阐述：

> 配额概率样本的基本假设是，被调查人是可以进行分层的，且层中宜于进行调查的概率是已知的……任何一个被调查人被调查的概率等于他最初被选到的概率与他的宜于进行调查的概率的积。（Sudman，1976，p. 193）

然而，配额概率抽样是一种有偏的抽样方法，尽管这种偏倚一般不算太大。此外，它的抽样误差也高于同样容量的其他概率样本（Hess，1985）。

非概率样本的用途

在某些场合非概率样本是一种有用和便于进行调查的抽样方法。在许多情况中，它是一种比较合适的抽样方法，而在某些时候，它是唯一可以使用的抽样方法。例如在某些特定人口总体，如非法使用毒品的人口总体的研究中，我们不可能得到抽取概率样本所需的清单。用于滚雪球式抽样的列举法可能是唯一可行的抽样方法。

在研究者的确对人口总体中的特定成员感兴趣的时候，用这些特定成员而非整个人口总体中的所有成员来构成样本时，采用非概率样本也许也比较恰当。例如，某些比较性的政府研究更感兴趣的是某些特定的国家，而非国家的分组（如发展中的债务国）。

在探索性研究中，研究者的目的在于确定某一问题是否存在，因此非概率的陈述可能是一种比较切合实际的选择。试点研究可以选择那些可能揭示问题的个案进行。由此得到的数据可用于考察问题是否存在。这样的方法虽然不允许我们估计问题的大小，但在资源紧缺的情况下，也许可以事先用它来有效地确定一个系统的调查是否的确需要进行。

虽然在许多情况下，使用非概率样本的必要性是不言而喻的，但是它的使用增加了使用样本数据来代表总体的不确定性。卡尔登对这一问题做了简明的概括：

> 非概率抽样涉及的各种各样的步骤和方法，包括志愿者的使用和在现场主观选取“代表”总体的样本单元等。所有的非概率抽样方法的通病是它的主观性。这种主观性使我们无法为它研发适当的理论框架。(Kalton, 1983, p. 7)

因为选择过程的主观性，在我们将非概率样本作为总体的代表时，不确定性会有所增加。混淆变量可能会对研究结果有所影响。有关总体陈述的正确性和精确性仅仅取决于主观的判断。与后面我们将要讨论的概率抽样不同，非概率抽样的选择的程序并没有为我们提供用样本结果推论总体的法则或方法。正因为如此，非概率样本都会存在着一种因选择过程的偏倚而导致研究结果无效的危险。

虽然使用立意样本的最大风险是外部效度问题，但是在研究发现的信度方面，它也同样也存在着一定的风险。我们常常会问这样一个问题：“如果我们为样本选择了其他的单元，那么结果又会是什么呢?”选择过程中不期而至的偏倚常常会使得到的结果与那些我们对总体期望的结果相去甚远。无偏的选择过程是保证我们能得到无偏样本的唯一方法。

斯图亚特指出：“样本本身决不会告诉它赖以产生的过程是否是无偏的。如果我们不希望永远也无法脱离选择的偏倚的阴影，我们就必须了解具体的选择过程是什么。”(Stuart, 1984, p. 4)正因为立意样本的结果的信度有赖于个人对选择方法和样本单元的判断，因而即使配额抽样得到的样本比例与总体的一致，它的选择也是有偏的——总是偏向总体中那些易于联系和调查到的个体。在这样的情况下，偏倚是由比例造成的。

总之，在有些情形下非概率抽样是我们得到数据的唯一途径。资源有限、无法确定总体成员、必须确定问题是否存在等都可以成为使用非概率样本的理由。然而，研究者必须清楚地意识到使用非概率样本带来的风险——对研究的发现的效度和信度带来的风险。不过在使用非概率样本之前，我们应该首先对本章下面一节讨论的概率样本，以及它的各种备择方法予以认真的考虑和全面的审视。

有关非概率样本误差最为著名的例子莫过于1948年的美国总统选举。当时三个最有名的调查公司使用配额样本进行调查，认为托马斯·杜威(Thomas Dewey)将会以很大的优势战胜哈里·杜鲁门

(Harry Truman)。而实际上杜鲁门得到了50%的选票,高于杜威的45%。尽管样本成员在地点、年龄、种族和经济地位上的比例与选民总体一致,但是调查员主观上偏向在共和党人中选取调查对象的偏倚,导致了预测的误差。在民意测验中的正确性和可信性上发生的意料之外的偏倚使得调查公司开始改而采用费用更高的概率样本。

概率抽样

概率样本的特点是能使总体中的每一单位都有一个已知的、非零的被包含在样本中的概率,尽管对于总体的所有单位而言,选择的概率并非总是相同的。给总体的每一成员以相同的选择概率的抽样设计叫做等概抽样。在总体的某些成员入选样本的可能大于其他成员时,我们得到的样本叫做不等概选择样本。为了弥补总体成员入选样本的可能性上存在的差别,研究者必须对不等概样本的数据进行修正。修正的方法是在估计时给某些个案以更大的影响力或权重。例如,在总体的某些成员入选样本的可能性四倍于其他成员时,那么在估计时它们的权重就必须减少为四分之一,或以 0.25 为权。缩小这些个案的影响将使这些个案出现在样本中的可能四倍于其他个案这一事实得到控制。基于选择概率的权重使我们得以用不等概的样本来无偏地代表总体。

在一个概率样本中,某些单位已确定要被选入,它们入选的概率为 1。这些单位是研究者主观断定样本中必须要有的个案。任何在伊利诺斯州地区的调查一般都会包括芝加哥的库克县(Chicago-Cook County)。在样本中芝加哥的库克县只代表它自己,为了弥补选择概率上存在的差异,在使用样本进行估计的时候,需要进行一定的修正。

概率样本意味着随机选择机制的使用。随机选择去除了选择过程的主观偏倚,从而使我们有了将样本结果推论到总体的理论依据。随机选择机制包括使用号码完全打乱的抽奖方法、从随机数码表中抽取一组数字,以及用计算机程序从一份自动生成的清单中生成随机的单位清单等。随机并非任意或偶然。随机选择是一种非常仔细的特别的程序,它能确保每一样本单位的选择都独立于其他单位的

选择。随机性转换成了每次选择的独立性,这就是说,总体中的任何一个成员的选择都不会对总体中任何其他正在被选择的成员的选择的可能性产生影响。真正的随机过程是难以做到的,且常常受到人们的主观错误的影响。麦克金恩在一篇题为《有序地追求纯粹的无序》(*The Orderly Pursuit of Pure Disorder*)的文章中以1969年的草案彩票(draft lottery)为例(根据1940年美国总统罗斯福签署的法令建立的独立的,司职平时和战时联邦军队兵员空缺的人员甄选的机构。草案彩票是该机构用于兵员甄选的一种特殊彩票,译者注),揭示了这种被普遍认为是随机的过程的灵敏性:

> 具体的操作步骤是,先将含有一月份生日的所有胶囊放进一个盒子并将它们搅和,然后再按照这一方法,依次将二到十二月份的胶囊放入盒中并搅和。这种做法导致那些装着月份较晚的胶囊的搅和程度低于月份较早的胶囊。这样在胶囊抽取的当晚,那些装有较晚月份的生日的胶囊首先被抽到:在最初三分之一的抽取中,装有十二月的胶囊的抽中的机会高于了平均数。有鉴于此,次年,为了避免这样的情况的发生,草案官员对放入程序进行了修正,以使不同月份的胶囊都能得到程度相同的搅和。(Mckean, 1987, p. 75)

选取概率样本的基本方法有下列5种。

- 简单随机抽样(Simple random sampling)
- 系统抽样(Systematic sampling)
- 分层抽样(Stratified sampling)
- 整群抽样(Cluster sampling)
- 多级抽样(Multistage sampling)

下面我们将对每一类样本做一个简要的定义。表2.2是每一类样本特点的概要,本书第6章将会介绍每一种抽样方法的必要条件、长处和不足。

表 2.2　概率样本设计

抽样类型	选样策略
简单随机	研究总体的每一成员都有相等的入选概率
系统	研究总体的每一成员既可以是在实地集合在一起,也可以是列在清单上的,先设计选定一个随机的起始点,然后以相等的间隔选取总体的每一成员
分层	研究总体的每一成员被分配到一个组或一个层,然后再在每一层选取简单随机样本
整群	研究总体的每一成员被分配到一个组或一个群,然后再随机地选取群,选出的群中的每一成员都将被包含在样本中
多级	先如整群抽样那样,抽取群,然后再用简单随机抽样法从选出的群中抽取样本成员,群的抽取可以分多个阶段进行

简单随机抽样。简单随机样本是以研究总体中的每一成员都有一个相等的被选取的概率这样一种方式选取的样本。研究总体中的所有成员既可以是物质的存在,也可以是清单上列出的名单。对总体中的成员进行不间断的随机抽取,直至抽够预先规定的成员或单位数为止。

一般我们以小写的英文字母"n"代表样本中的单位数,而以大写的英文字母"N"代表研究总体中的单位数。简单随机样本具有这样一个特性,那就是每一个含 n 个单位的子集都有相等的从含 N 个单位的总体中被抽到的可能。研究总体中每一个成员被选中的概率是:

$$p = f = \frac{n}{N}$$

式中,p 是选中的概率,

f 是抽样分数。

例如在一个有关高龄老人研究中,我们从一个含 625 000 个成员的总体中选取了 1 600 个 75 岁以上的成年人。个体被选中的概率是 1 600/625 000,或 0.002 6。

本书使用的简单随机选取假定选取是无回置的。这就是说,在

抽样过程中,一个单位一旦被选中,它将从做进一步抽取的合格成员库中去除。这与回置选取是不同的。在回置选取中,选中的单位将被放回做进一步选取的合格成员库。尽管回置选取具有我们所希望的统计学性质,但是它却可能在操作上有诸多不便,因为它有可能使某一特定的单位被多次抽中。

系统抽样。有时我们也把系统样本叫做“伪简单随机样本”,因为它们有与简单随机样本类似的性质,而在有些实际场合,这种方法更为简单易行。系统样本的选取过程是:首先设法将总体成员集合起来或制作总体的清单,再选定一个随机的起始点,然后选取每间隔 i 个位置上的单位(如第 6 或第 234)。随机的起始点是这一过程的要点。没有随机的起始点,研究总体中的某些成员的选中的概率就可能为零。这样样本就不能被视为概率样本。

选择间隔 i 取决于公式:

$$i = \frac{N}{n} = \frac{1}{f}$$

为了实地操作简便起见,间隔必须是一个整数。在商为一个分数时,我们推荐使用的取整方法是向下取整,即将 12.895 向下取整为 12。因为我们采用了向下取整,便会有多于 n 个的单位被选中。然后再通过第二次随机选取,将超过我们要求的 n 个单位的那些单位去除。例如,如果我们希望的样本量是 1 000,并将选取间隔 12.895 取整为 12,那么我们预期可以产生1 075个左右的样本个案。随后我们可以采用简单随机抽样法,在这 1 075 个个案中选取 75 个个案,在数据收集之前将它们从样本中去除。

如果总体清单的排列具有某种周期性的方式,且周期又与选取的间隔一致,那么就可能产生某种问题。如果我们选择了 12 作为选择的间隔,而数据又是以月为次序排列的,那么将会从每一年中选中同一个月。在采用系统抽样法时,我们应该避免对总体进行周期性的排列,或将单位重新排列。

分层抽样。分层抽样要求在抽样之前将调查总体的成员分为组,这样的组也叫做层。根据我们事先对单位的了解,将每一个单位分配到某一层,且只能分配到一个层。然后,用类似前面在简单随机

抽样或系统抽样中介绍的两种抽样方法中的某一种,从每一层抽取独立的随机样本。

分层抽样可以以相同的抽样分数这样一种方式进行,即将 f 用于每一层的抽样。这样的抽样叫做成比例的分层抽样,它是一种等概选择法。有时,我们也可以不同的比率为每一层设定抽样分数。采用不同抽样分数将导致选择的不等概,而这样的分层抽样则叫做不成比例的分层抽样。不成比例的分层抽样需要研究者做一些额外的工作(如在分析时给答案加权),但是这种额外的工作将会因分析的精度有所提高而得到回报。

在组的代表性对于研究或政策的制定是很重要的时候,为了确保组的代表性,这时我们通常都应该使用成比例的分层抽样。例如,一个全州范围的高龄老人研究,如果州的每一个地区的代表性是很重要的,那么我们就应该采用成比例分层抽样。不成比例的分层抽样的使用,通常都是为了能对某些特定的层的成员进行分析,或提高整个样本估计值的精确度。不成比例的分层抽样的具体策略和实施要点我们将在第 6 章介绍。

整群抽样。整群抽样表面上看来与在抽样之前将总体成员分成唯一的、不相重叠的分层抽样颇为相似。在整群抽样中,我们将组叫做群,包含在样本中的群一般都是自然形成的,如学校、住户或诸如城市的街区这样的地理单位等。而我们随机地选取的是群,选取的群中的每一个成员都将包含在样本中。在整群抽样中,抽样单位是群,而非总体中的单独的成员。

整群抽样与分层抽样不同,整群样本只涉及少数几个组的选择,数据将从选中的组中的所有成员中收集,而分层抽样则要从每一组或层选取少数几个成员。我们用一个有关学校的例子对此做进一步的说明。为了从一个学校选取学生,每一个学生可以被置于一个以年级区分的层。然后我们可以从每一层随机地选取若干学生得到一个分层的样本。我们也可选择另一种方法,先将学生按教室分为群,然后我们可以选取一个教室的样本,并将选中的教室中的所有学生作为样本。

随机地选取群这一要求,使整群抽样法成为了一种概率抽样方法。表2.3列出了整群样本的选择概率的计算方法。该表数据根据

一个分为 100 000 个群(C)的 5 000 000(N)个住户单位的总体计算的。在该例中我们随机地选取了 40 个群(c)。这一例子中的群也可以是城市中的普查片(census tract)。因为群的容量为 50 左右,且选中的群中的所有住户都被包含在了样本中,所以样本的容量(n)等于 2 000(50×40)。整个的选择概率等于 0.000 4 (2 000/5 000 000 = 0.000 4),尽管各个群的住户数不尽相同。

表 2.3 整群抽样的选择概率

总体信息
总体成员数:$N = 5\ 000\ 000$
总体群数:$C = 100\ 000$
群平均容量 = 50
样本信息
样本成员数:$n = 2\ 000$
样本群数:$c = 2\ 000/50 = 40$
概率:$p = 40/100\ 000 = (50 \times 40)/\ 5\ 000\ 000 = 0.000\ 4$

虽然分层抽样的使用一般都会使统计值的精确度有所提高,然而整群抽样的使用结果却恰好相反。整群抽样一般都会使统计值的精确度有所下降。整群抽样的使用一般都出于实际操作的考虑。在没有整个总体的清单时,我们可以因地制宜地抽取整群样本。在采用个别访谈作为收集数据的方法时,我们也可以抽取整群样本,从而大大降低交通和培训的费用。

多级抽样。在多级抽样中,我们使用与构成整群抽样相同的概念,只是把它扩展到多级抽样而已。最简单的多级抽样是二级抽样。在第一级,我们将研究总体的成员分组,先选取组。这些组叫做初级抽样单位(PSUs),类似于群。在第二级,则在前面选出的初级单位中,随机地选取总体成员。

选择的概率取决于累计的选择概率。因此为了使整个选择的概率相等,某一级中的不等概可以在随后各级得到弥补。表 2.4 列出了两种可能的情况。在表 2.4a 中,我们先等概地选取了 500 个初级

抽样单位，然后再从每一个初级抽样单位中选取 4 户。因为初级抽样单位含有的住户数是不同的，所以选择的结果是不等概的。在这张表中，我们可以看到初级抽样单位含有的户数有 50 户和 100 户两种。那些选自含 50 个单位的初级抽样单位的个案的选择概率为 0.000 4，而选自含有 100 个单位的初级抽样单位的个案则有选择概率 0.000 2。表 2.4b 中的例子表示概率与大小成比例(PPS)的选择方法。这时，初级抽样单位是根据自身的大小选取的：

$$c \times N_c / N = p$$

式中，c 为选取的 PSU 的数目，

N_c 是在一个特定的 PSU 中的元素数，

N 是总的元素数，

p 是选择的概率。

表 2.4　多级样本的选择概率

总体成员数：$N = 5\,000\,000$

初级样本单位数：PSU = 100 000

样本容量：$n = 2\,000$

不等概选择的样本

a PSU 中的单位数	第一级 等概选择 500 个 PSUs	×	第二级 每个 PSU 选择 4 个单位		选择的概率
50	500/100 000	×	4/50	=	0.000 4
100	500/100 000	×	4/100	=	0.000 2

等概选择的样本：概率与大小成比例

b PSU 中的单位数	第一级 以与大小成比例的概率选择 100 个 PSUs (PPS)	×	第二级 每个 PSU 选择 20 个单位		选择的概率
50	(100 × 50)/500 000	×	20/50	=	0.000 4
100	(100 × 100)/500 000	×	20/100	=	0.000 4

在本例中，含有 50 个个案的 PSU 的选择概率是 100 × 50/

5 000 000，或0.001。含有100个个案的初级抽样单位被选中的可能性是含50个个案的两倍($p=0.002$)。然而通过从所有的初级抽样单位中选取数目相等的个案——在本例中是20，最终的样本单位的选择概率都是相等的($p=0.000\ 4$)。作为验证，总的选择概率乘以总体单位数，应该等于样本的容量($0.000\ 4\times5\ 000\ 000=2\ 000$)。

在采用多级抽样法时，我们需要在样本估计值精度的提高和降低数据收集的费用，以及因此而可能导致的使抽样过程更为复杂这几个问题之间进行反复权衡。一般讲，随着级数的增加，精确度也会有所提高。我们可以在任何，或所有的阶段引进分层的方法。分层方法的引进一般都会使精确度有所提高，但是费用和复杂性却会因此而有所增加。由密歇根大学调查研究中心(the Survey Research Center at the University of Michigan)实施的全美家庭调查的抽样，共使用了五级。分级和分层的问题将在本书第4章进行讨论。

结　论

抽样方法可分为两类，概率抽样和非概率抽样。在两种类型的方法之间进行选择时，必须从实际出发，在对每种备择方法所需的时间和人力物力等有所估计的前提下，考虑调查要求达到的效度和信度。设计精良和执行规范的概率抽样较之非概率抽样有更高的效度和信度。尽管情况也许并非总是如此，但在大多数情况下，抽取一个概率样本所需的费用总是高于非概率样本。

本书余下的篇幅将主要讨论概率抽样。尽管在某些场合，非概率抽样是很有用的，但本书的基本观点是概率抽样无疑是备择的抽样方法的首选。换言之，只有在概率样本无法使用的时候，非概率抽样才在我们考虑之列。

下一章我们将把关注的重点放在概率抽样之所以成为备择抽样方法之首选的基本原理上面。我们可以对因概率抽样的使用而造成的偏倚和可能的误差做严格的审视和估计，而我们无法对非概率样本进行类似的审视和估计。因为，由于某一个样本的使用而导致的不确定的范围可以用为概率样本设定的置信度来进行估计，但非概率样本却无法做这样的估计。已经充分发展并在概率抽样中得到检

验的理论可以对这种差别作出解释。

第 3 章和随后的几章将更加详细地为研究者介绍一些实用的抽样设计方案和用于设计和实施概率抽样的框架。如果有相应的研究目标和可资利用的资源,这些框架和实例应该在构建实用的设计方案时会对研究者很有帮助。为了能最大可能地得到适合自己的研究的样本,理解和掌握这些工具尤其重要。

第3章

实用样本设计法

Practical Sample Design

实用样本设计法的目的在于得到符合我们的研究所要求的有效和可信的数据和统计值。实用样本设计法是一种把抽样设计和实施看做整个研究过程的一部分,并用总误差这一概念来判断效度、信度和精确度的方法。数据的效度会对用样本结果推论总体的精确性和从假设检验中得到的结论的正确性有影响。在大规模的测量中,信度取决于样本选择的过程。采用随机选择法,能使我们避免样本选择过程中的主观随意性,从而使信度能有所提高。

对于精度问题的关注是使用抽样方法的必然结果。选取总体的某一子集意味着总体的某些成员未能被包含在样本中。正因为如此,用同样的程序重复地选取样本得到的结果将会有所不同。令我们感到欣慰的是,抽样理论为我们提供了样本结果波动的十分有用的信息。借助抽样理论我们可以计算可能的波动量和抽样的变异性。抽样理论已经非常清楚地刻画出了抽样的变异性和精确性这二者之间的关系:样本统计值的精确度随样本的变异性的增加而下降。抽样理论使研究者得以了解导致变异性的因素究竟是哪些,进而找到降低变异性的途径,使数据达到要求的精确度。

为了理解在设计过程中我们究竟应该如何做出抉择和使用实用样本设计法,我们必须首先了解可能对效度和抽样的变异性构成威胁的因素有哪些。在第1章介绍的那些对外部效度构成威胁的因素,在抽样文献中常被叫做偏倚。样本选择的偏倚将导致样本和样本所代表的总体之间出现系统差异。在前面的章节中提到的偏倚源包括目标总体和研究总体之间存在的差异,以及因抽样方法使用不当而使样本中的某些子群体的数目超过比例等。在抽样的变异性增

加时，由此而造成的对统计结论的效度的影响通常是显而易见的。抽样变异性的增加虽然并不会引起总体和样本之间的系统差异，但是它却会对估计值的精确度有所影响，进而可能导致我们无法得到正确的结论。偏倚和抽样的变异性二者一起构成了样本总误差。在下一章，我们将对总误差进行系统的分解和分析。本章的最后一节将介绍实用样本设计的框架，它将在可能的范围内为我们提供进行抽样抉择的指南，以降低偏倚和抽样变异性。

抽样设计中的误差源

实用抽样设计的目的在于在研究者的目标和可资利用的资源既定的前提下，最大限度地降低样本选择过程的总误差量，使样本达到要求的精确度。总误差由三个独立的部分组成：

- 非抽样偏倚：如因为总体定义的差异或测量误差引起的系统误差。
- 抽样偏倚：由抽样方法不当导致研究总体某一部分超出比例而引起的系统误差。
- 抽样变异性：由随机选择过程引起的样本估计值围绕研究总体参数值的波动。

尽管总误差的每一种组成部分都应该引起研究者的注意，但是在抽样设计的时候，尤其要注意的是后两种误差。在多数情况下，样本误差的讨论都集中在抽样的变异性和它的估计值、标准误差或抽样误差等问题。对于有效的抽样设计而言，误差的所有三个来源都应该在认真的考虑之列。

我们将总误差定义为目标总体的真总体值与基于总体的样本估计值之间的差异。均值的总误差是：

$$E = \overline{X}_T - \overline{x}$$

式中，X_T表示真的总体值；

$\overline{x}$ 表示样本均值。

因为总误差会对研究目标的实现有影响，所以实用样本设计必

须通盘考虑总误差的三个组成部分。总误差的三个组成部分中的每一个部分在研究过程中有可能发生的时间及每种误差源的某些实例都在图3.1中做了形象的阐述。不仅如此,我们都是在资源有限的前提下来考虑样本的设计问题的。为了降低某一类误差而做出的资源分配的决策会对降低另两类误差所需的资源有影响。在实用抽样设计中,重要的问题是权衡利弊,通盘考虑。可用于设计的资源总是有限的,它令研究者不得不在如何降低总误差的各个部分时要反复进行考虑,权衡各方面的利弊得失。研究者必须充分意识到,有关降低三类误差的决策,都必须基于有利于降低总误差这样一个前提。

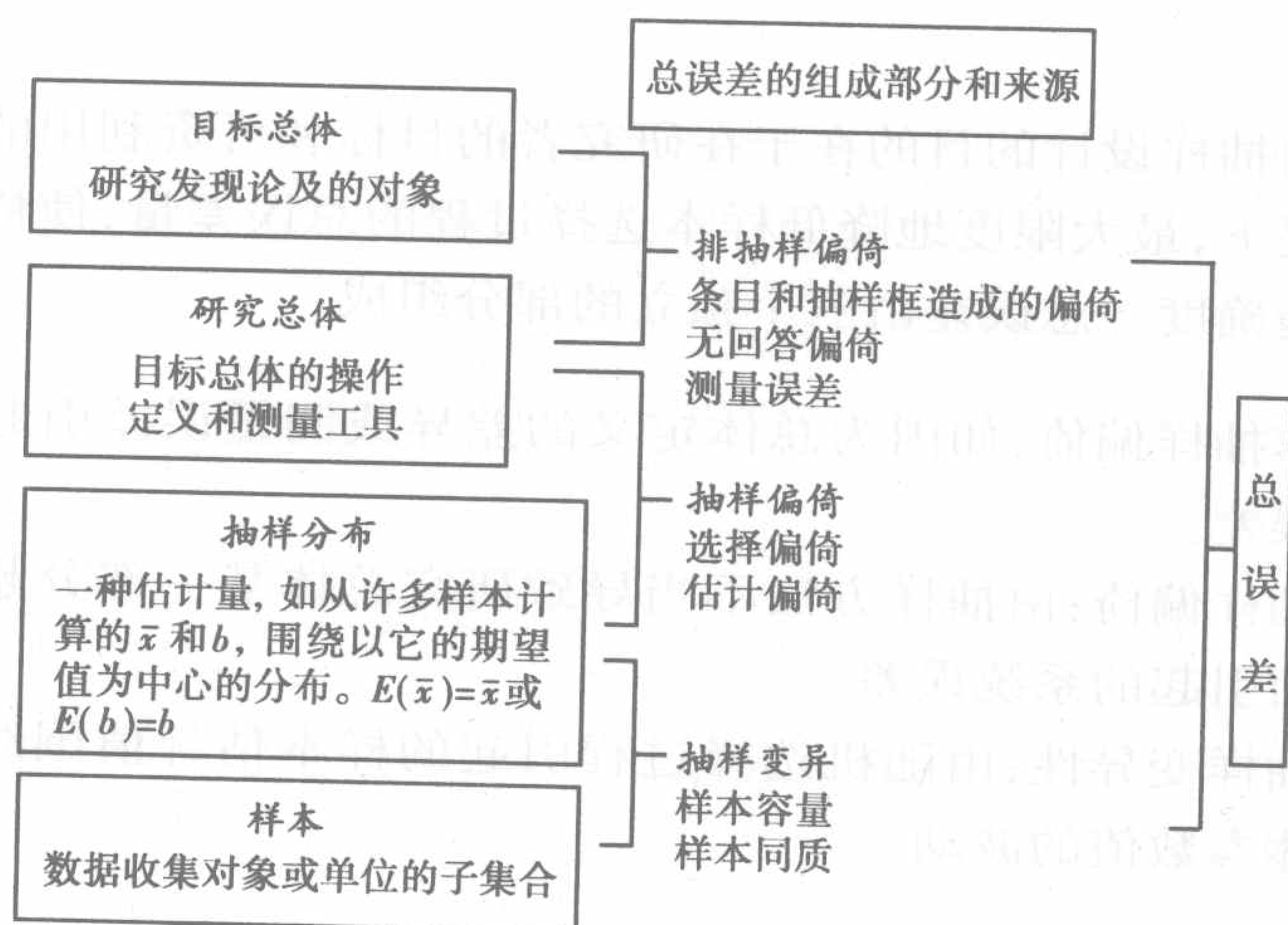


图3.1 实用抽样设计

非抽样偏倚。如果数据收集是实施于整个总体的话,那么非抽样偏倚便是真实的目标总体值与我们得到的总体值之间存在的差异,非抽样偏倚由具体的决策及在数据收集过程中决策的具体实施引起的,它们和样本选择并没有什么直接的关系。例如,研究总体的定义可能将目标总体的某些成员排除在外,但这些成员却是研究者希望能包含在自己的研究发现中的。在这样的情况下,即使我们收集了整个研究总体的数据,最终的发现也会因为目标总体的某些成员被排除在外而有所偏倚。以总体的均值为例,非抽样偏倚的计算公式是:

$$NSB = \overline{X}_T - \overline{X}_S$$

式中，$\overline{X}_T$ 表示真实的目标总体均值；

$\overline{X}_S$ 表示从研究总体得到的均值。

有若干因素可能会引起真实的总体均值与调查总体均值之间出现差异，其中与样本设计有关的主要差异是目标总体和研究总体之间存在的差异。所谓目标总体就是研究者要探讨研究的群体。目标总体的定义既可以以将要检验的理论的要点和关注点为依据，也可以以来自将要考察的政策的关注点为依据。例如，在对无家可归者的需求进行全面评估时，目标总体应该包括所有的无家可归者，无论他们当时是否已经得益于已经实施的项目。另一方面，在对提供给无家可归者的社区精神健康服务进行评估时，则只应该包括那些接受社区精神健康治疗的无家可归者。较之社区精神健康服务评价的目标总体定义，需求评估目标总体的定义更为宽泛，包括了所有的无家可归者，而后者只涉及了其中的一部分。目标总体的定义必须服从于研究目的。

研究总体将目标总体操作化。目标总体大多是动态的，因为不时会有新成员加入和老成员离去。已经制成的目标总体清单可能是不完整的，而某些成员可能也是难以确认的。在另一些时候，研究总体中包含了某些并不属于目标总体的成员。例如，也许有的研究者对 4 岁儿童的发展项目感兴趣。在一个用于项目评价的探索性研究中，研究总体可能包含了 3.5 到 5.5 岁的儿童。尽管研究中包含了年龄大于或小于 4 岁的儿童，但研究者仍然想把从中得到的发现推论到 4 岁的儿童。因为发展项目规定的年龄可能会对结果有一定影响，所以将年龄组扩大可能会产生某种问题。然而，如果探索性项目能采用与项目在大规模实施时相同的定义，那么在这一例子中发生的因年龄组的扩大而造成的问题就可以减少。为了确保探索性研究与构成项目实施基础的理论和项目资助者的本意相一致，我们必须从项目一开始就考虑和讨论这一问题。

除了定义问题之外，其他能引起目标总体与研究总体差异的因素还有无回答偏倚。无回答偏倚是由于无法与总体的某些成员联系，或某些成员拒绝提供调查所要求的数据造成的。在发展项目这一例子中，在认知试验进行的时候，试验的班级中的缺席者便可能引

起无回答问题。如果无回答完全是随机的，它便不会产生偏倚。但在许多时候情况并非如此。在更多的时候，无回答者都来自总体的某个可以比较确定的子群体，而在实际收集的数据中，来自那一子群体的数据的缺失将导致结果发生偏倚。

研究者决不可以简单地假设无回答是无偏的。在通常情况下，由研究者采用的程序引起的（如只在白天，而不是同样在晚上也与总体成员进行联系）无回答，或无回答中存在某种系统的模式都会造成无回答者在样本中比例不够。处理无回答问题偏倚的最好方法是减少无回答的存在，从而减少总体中代表性不足部分的比例（Kalton，1983，p. 64）。锲而不舍地对起初没有成功的调查进行追踪，尝试用多种方法与样本成员进行联系，以及采用那些能最大程度地减少被调查人拒绝参与调查的方法是降低无回答数量的最为切实可行的手段。

其他导致非抽样偏倚的因素还有测量误差和那些在记录、编码和数据转换过程中发生的误差。这两个问题超出了实用抽样设计讨论的范围，尽管它们对于降低总误差十分重要（参见 Raj，1972）。测量误差的含义在库克和坎贝尔（Cook & Campbell，1979）的有关著作中进行了介绍。布拉德本和苏德曼（Bradburn and Sudrman，1980）通过对题项用词的仔细斟酌，使我们得以从经验研究的角度深入了解如何在调查研究中降低回答偏倚。福勒（Fowler，1984）的著作为我们如何在各种抽样调查研究中减少两种来源的偏倚提出了许多有益的建议。拉夫拉卡斯（Lavrakas，1986）则专门就如何在电话调查中减少两种来源的误差提出了很多建议。

抽样偏倚。抽样偏倚是在研究总体的值与期望值之间存在的差异：

$$SB = \overline{X}_S - E(\overline{x})$$

式中，SB 表示偏倚；

$E(\overline{x})$ 是均值的期望值。

均值的期望值是对研究总体反复进行抽样得到的均值的平均值。均值的期望值等于研究总体的均值，如果抽样和计算程序是无偏的话。

抽样偏倚可被分解两个部分：（1）选择偏倚；（2）估计偏倚。在

研究总体的所有成员的选择概率不相等的时候，便会因此而引起选择偏倚。我们可以用估计程序对不等概进行修正。修正的方法是通过加权对选择的不等概作补偿。

从一张含有某些成员的重复条目的研究总体的清单中进行抽样便是一个有关选择偏倚问题的直观的例子。在第 4 章介绍的那个市民调查的例子中，研究总体的清单是由两张清单合并而成的：州纳税申报单和享受医疗补助的名单。如果某一个人同时出现在两张清单上，那么他入选样本的可能将会加倍。但从实际的操作角度看，将重复条目从合并后的清单中清除出去是不可能的。不过我们却可以对在选择过程中出现选择的不等概性作修正。

为了修正这种选择的不等概，我们可以在估计过程中，用一个等于选择概率增加量的倒数的权(w)：

$$w = \frac{1}{p} = \frac{1}{2} = 0.5$$

因为这种类型的个人的选择概率是出现在研究总体清单中一次的人的两倍，所以为了弥补他们出现在样本中的可能性的增加，他们应该得到一个为其他成员的一半的权。

在将某种估计法用于来自某一总体的所有可能的简单随机样本，计算得到的平均数不等于研究总体的值时，估计偏倚就会因此而产生。例如，计算得到的中位数是总体均值的有偏估计值。在估计时，为了对选择的偏倚进行补救，要进行种种调整，这种做法必然会将选择偏倚和估计偏倚联系起来。

统计量的期望值这一概念不仅加强了统计学理论，而且由于它在现代抽样实践中的使用，为许多切实可行的解决方法提供了依据。而均值这一估计量则为我们提供了更加深入地考察这一概念的便于使用的实例，虽然我们也可以使用许多其他的估计量，如标准差或回归系数，但它们都不如均值方便。均值的期望值是从研究总体的反复抽取的样本中计算得到的均值的平均值。从每一个样本计算得到的均值形成了一种围绕均值的期望值的分布，我们把这种分布称为抽样分布(图 3.2)。

当样本量达到 30 或更多的时候，抽样分布的形状就成为了正态分布，即形成一条如图 3.2 所示的形似一个钟的曲线。我们可以证明，不

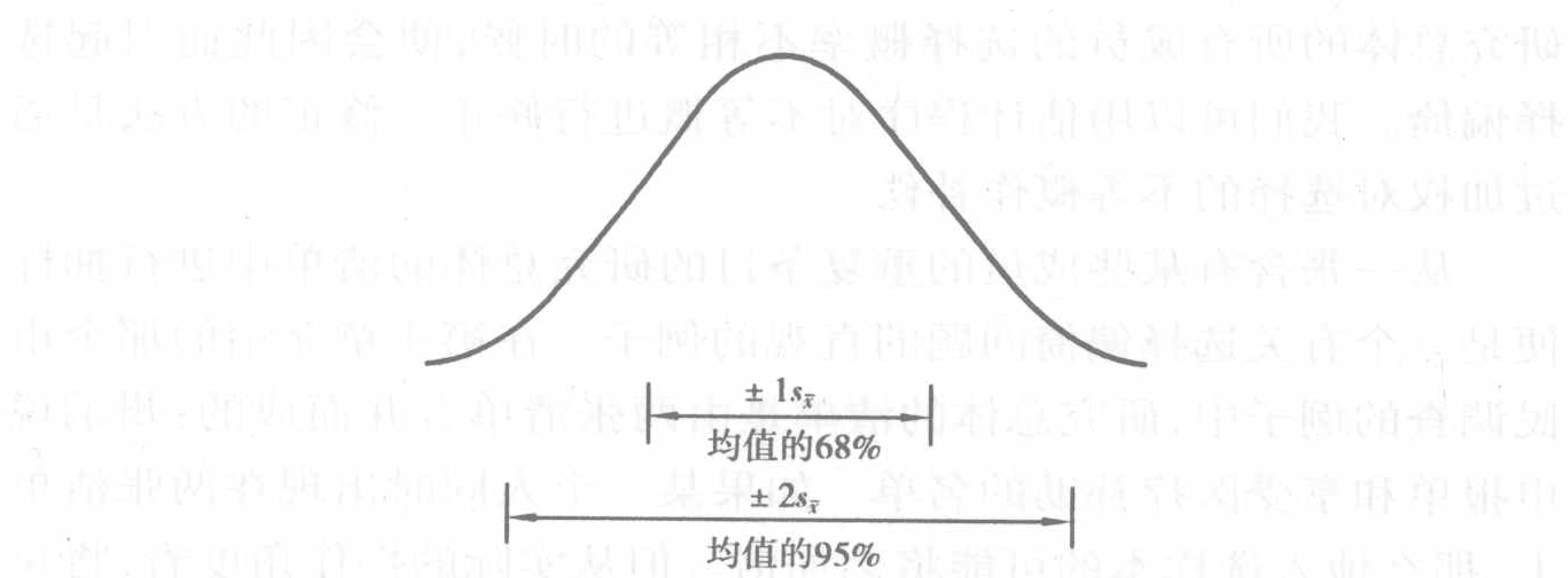


图 3.2 均值的抽样分布

论研究总体的频数分布的形状如何，只要选择的样本数足够多，抽样分布的形状都是正态的。

除了均值（期望值）之外，抽样分布的标准差则是这一分布的另一个重要的性质。它的正态分布的特性使我们能计算在研究总体的均值的某一确定数量的标准差单位范围内的样本均值的比例。图 3.2对抽样分布、抽样分布的均值，以及在样本均值的平均数一或两个单位的标准差范围的均值的百分比做了直观的阐明。

统计理论还告诉我们，抽样分布的标准差（$s_{\bar{x}}$）与样本大小反相关。这就是说样本量越大，抽样的标准差越小。图 3.3 用图形给我们展现了这种关系。

表示正态曲线下的面积的那些表格可以用于计算样本均值落入特定的单位标准差数的百分比。对于比较小的样本，我们可以用学生 t 分布。在样本量小于 30 时，那些值（标准差单位数）会因为样本量较小而变得比较大。例如，在样本量为 100 时，将有 95% 的均值落入 ±1.96 个标准单位这一范围内。在样本量为 10 时（9 个自由度），包含 95% 的样本均值需要 ±2.26 个标准差单位。当样本量下降到低于 30 个左右的单位时，为了能包含一个特定的样本均值的比例，所需的标准差单位数必定会有所增加。正因为如此，小样本的置信区间相对较大。大多数列有 t 分布的表格中都可以找到这些 t 值。首先我们必须计算自由度（df），方法是将样本容量减 1（$n-1$）。在这一例子中，样本容量为 10，所以自由度应该等于 9。然后，选择与 1 减去概率水平的差值对应的那一列。这就是说，在概率为 95% 的时候，要用 $\alpha = 0.05$ 那一列。选择的列是用于双尾检验的。为了确信

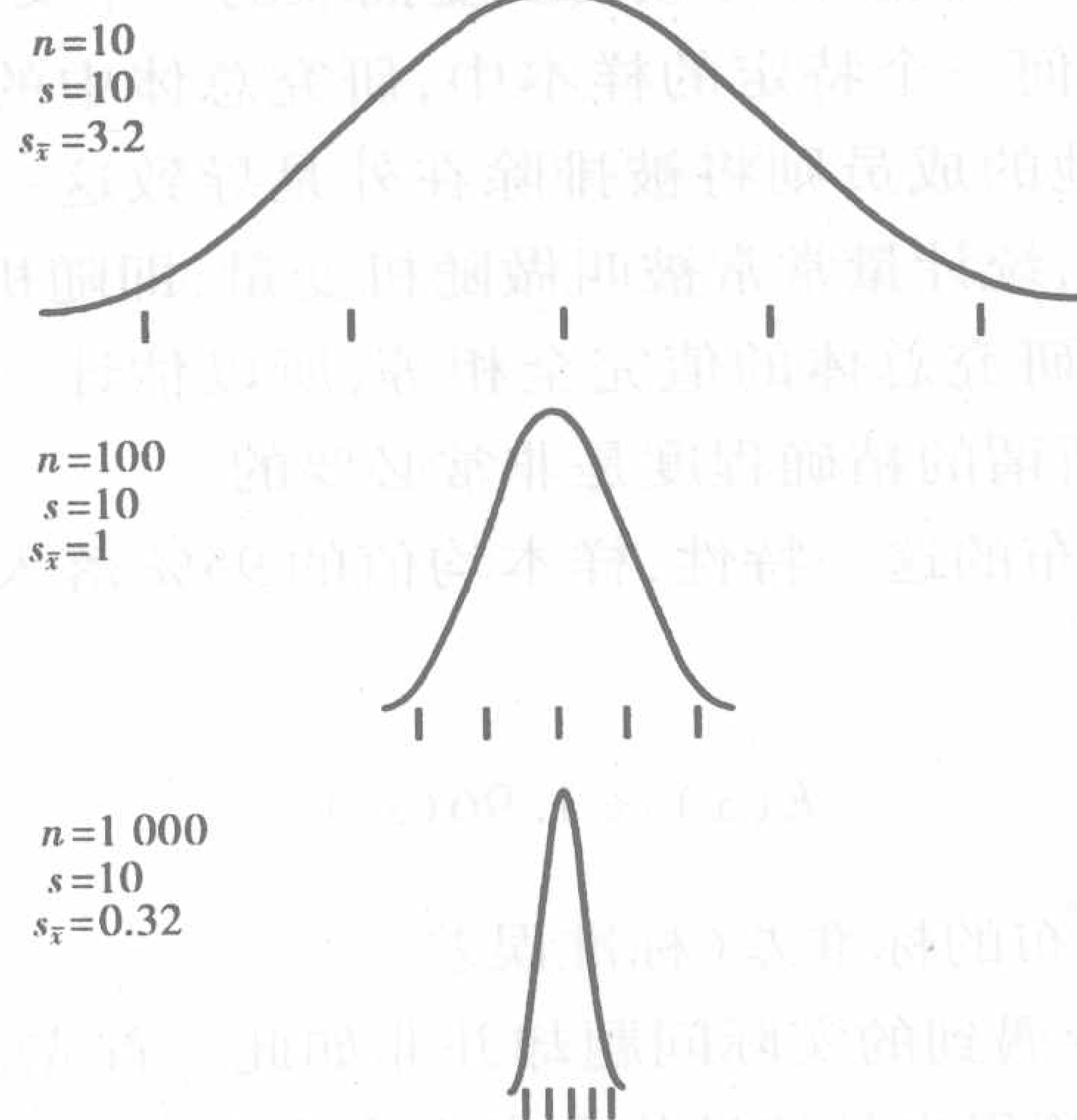


图 3.3 备择样本容量抽样分布比较

这一点,我们不妨将这一问题与我们使用的表格的方向联系起来考虑。如果表格显示的是单尾检验的数值,或只有一侧的图形有阴影,我们需要将 α 的值加倍。例如,一张显示单尾检验表格,α = 0.025 将用于95%的置信度。

最后,再找出表中位于含有合适的自由度的行和选定的概率水平的交叉点上的那个单元格。一个容量为10的样本,置信度为95%的那一单元格的值为2.26,具体如下面所示:

df	α = 0.05	α = 0.01
1	12.71	63.66
⋮	⋮	⋮
9	2.26	3.25
10	2.23	3.17

这就是说在均值两侧的2.26个标准差单位这一范围内,将含有95%的样本均值。若置信度为99%,则需要3.25个标准差单位。

抽样的变异性。样本总误差的最后一个组成部分直接由这样一

个事实所致,即由于偶然性导致的反复抽取的一个又一个样本,彼此不尽相同。在任何一个特定的样本中,研究总体中的某些成员将被包含在内,而其他的成员则将被排除在外是导致这一变异的主要原因。正因为如此,统计量常常被叫做随机变量,即随机而变。因为统计值通常并不与研究总体的值完全相等,所以估计一下它对总体值的接近程度,即所谓的精确程度是非常必要的。

使用抽样分布的这一特性,样本均值的95%落入的区间是可以计算的:

$$E(\bar{x}) \pm 1.96(s_{\bar{x}})$$

式中,$s_{\bar{x}}$是抽样分布的标准差(标准误差)。

然而,研究者遇到的实际问题却并非如此。首先,研究者希望了解从样本数据计算得来的统计值究竟在多大程度上接近总体值。其次,抽样分布的标准差在实际研究中几乎都是未知的,因为样本只有一个,而不是反复抽取的许多个样本。

第一个问题很快就可以解决。解决的办法是利用上面提到的抽样分布的性质。即使用它表明的抽样分布的均值落入与之关联的抽样分布的均值的标准差的单位数的百分比(如95%的均值将落入$\pm 2s_{\bar{x}}$),这一问题便迎刃而解。在计算出样本均值之后,我们可以围绕这一值来绘制一个区间,诸如这样的区间将有一个包含研究总体的均值预定概率:

$$\bar{x} \pm t(s_{\bar{x}})$$

式中,t是刻画预定概率水平的t统计量;

$s_{\bar{x}}$是均值的标准误差。

在本例中,围绕样本均值绘制的区间是研究者所期望的真均值能够落入的那个区间。与这一期望有关联的置信度以选择的统计量的预定概率为依据。这一区间常被叫做置信区间。

研究者将会遇到的第二个问题是抽样分布的标准差或标准误差的估计。在解决这一问题时,我们必定会用到总体的标准差的性质、总体标准差的样本估计值和抽样分布的标准差。总体的标准差是总体均值的离差的平方和除以总体含有的单位数的商的平方根:

$$S = \left[\frac{\sum (x_i - \bar{x})^2}{N} \right]^{1/2}$$

我们可以用下面的公式，即用样本数据来估计总体的标准差：

$$s = \left[\frac{\sum (x_i - \bar{x})^2}{n - 1} \right]^{1/2}$$

抽样分布的标准差是每一样本均值与所有样本的均值的平均数的离差的平方和除以样本均值个数的平方根：

$$S_{\bar{x}} = \left[\frac{\sum (\bar{x} - \bar{X})^2}{m} \right]^{1/2}$$

式中，m = 样本均值的个数。

这一统计量称作标准误差或抽样误差。

抽样理论告诉我们，抽样分布的标准差通过下面的公式与总体标准差的样本估计值相关联：

$$s_{\bar{x}} = \frac{s}{n^{1/2}}$$

由该公式可知，对抽样的变异性有影响的因素有两个：变量的变异性（标准差）和样本容量。比较小的标准差会使均值的抽样误差有所降低。样本容量越大，抽样分布的标准差越小。

因为总体的标准差可以用样本数据来估计，而样本的容量是已知的，所以我们可以用一个公式来估计抽样分布的标准差，以后我们将称之为估计值的标准误差。在这一特定的例子中，它就是均值的标准误差：

$$S_{\bar{x}} = \frac{s}{n^{1/2}}$$

$$s = \left[\frac{\sum (x_i - \bar{x})^2}{n - 1} \right]^{1/2}$$

式中，$s_{\bar{x}}$ 是标准误差的估计值；

s 是标准差的估计值；

n 是样本容量；

x_i 是样本观察值；

$\bar{x}$ 是样本均值。

研究者可以用这一公式来估计标准误差的值。估计得到的这一统计值将用于测量总误差的最后一个组成部分。这一估计仅以来自样本的信息为依据。

本书讨论的概率抽样设计假定样本是无回置地被选取的，这就是说，一个总体单位一旦被随机地抽入样本，它将被置于一边，因而不具有再一次被抽取的资格。随着更多的个体从总体被抽出，无回置抽样将使总体中可以被选择的个案数受到限制。由此造成的总体的有限性，可能需要在计算估计值的标准误差的时候，引进有限总体的修正因子（FPC）。

均值的标准误差的使用 FPC 的公式是：

$$s_{\bar{x}} = \frac{(1 - n/N)^{1/2} s}{n^{1/2}}$$

大家应该记住这样一条经验法则：在使用 FPC 时，样本量必须在总体容量的 5% 以上。这是因为，在抽样分数小于 0.05 的时候，纠正因子非常接近 1，它将会对标准误差的计算产生不恰当的影响。

标准误差的计算方法都是针对我们所估计的特定的统计值的。例如，我们普遍用于比例的标准误差的公式是：

$$s_p = \left(\frac{pq}{n}\right)^{1/2}$$

大多数统计教科书都会介绍若干种估计量的标准误差的计算公式。同样，用于计算这些统计值的公式也几乎为所有的统计软件包所使用。这些公式，譬如上面介绍的那些公式，都假定样本选择采用的是简单随机样本设计。用于更为复杂的抽样方法的公式将在第 7 章介绍。

在术语问题上我们必须进一步提醒读者的是：抽样误差和标准误差常在文献中交替使用。它们是用于测量抽样变异性这一更为普遍的概念的特定的统计量。不过，就二者而言，标准误差是更为恰当的术语。然后遗憾的是由于两个原因导致抽样误差这一术语使用更为普

遍。第一个原因是它含有误差是由程序引起的，而非自然发生的这一意义。第二个原因是它经常取代含义更为广泛的总误差这一概念。计算标准误差的实例将在第 4 章中予以介绍。

总误差。抽样设计是一种为了最大限度地降低总误差的三个组成部分，而有意识地进行平衡比较的过程。然而遗憾的是，人们却常常顾此失彼，只把降低标准误差变成了样本设计的唯一的关注点，因为它是可以被估计的。由于另两种偏倚成分不是很容易计算的，所以在设计过程中，它们常常未能引起足够的关注。然而，未能给总误差的所有三个组成部分都给予足够的关注并使它们减少，将会导致研究发现的效度和信度的降低。

我们用图 3.4 中的图像对总误差这一概念做了形象的概括。图像的顶端显示了目标总体的分布频数，它以完整的信息为依据。这一分布包含了总体每一个成员的值，排列的程序是由低到高（从左到右）。非抽样偏倚用目标总体的真均值与研究总体的观察均值二者之间的差表示。非抽样偏倚既包括由总体定义引起的差异，也包括由工具问题产生的误差和实际操作产生的误差所引起的差异。

我们用研究总体的观察均值与抽样分布的均值（$E(\bar{x})$）之差对抽样偏倚做了形象的阐述。而这些差异是由选择的偏倚或估计过程的偏倚造成的。最后，我们又对抽样变异性做了形象的解释。在本例中，它是抽样分布的均值和它的样本估计值（$\bar{x}$）之间的差。

我们将在以下一节介绍实用样本设计的框架，它将对研究者在考虑总误差问题时有所帮助，进而为他们在设计过程中需要决策的时候，提供可以依据的标准。

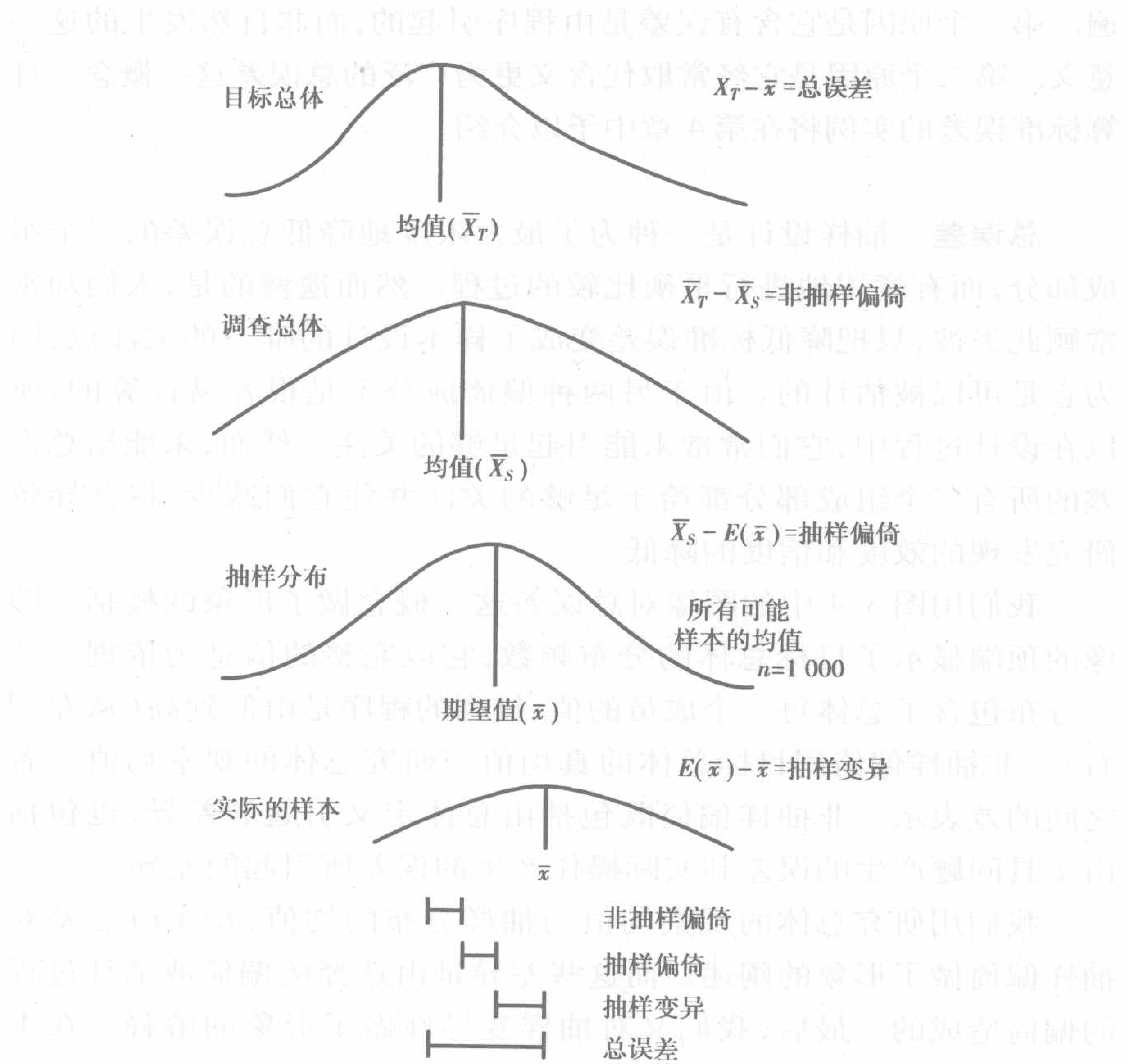


图 3.4　样本设计的总误差

实用抽样设计的框架

实用抽样设计的框架是一种在样本设计中用来启发研究者思维的工具。框架实际上无非就是一系列必须作出的抉择，而每一种抉择对整个研究都是不可或缺的。提供框架的目的在于帮助研究者和研究结果的使用者在考虑问题时，能围绕设计所需作出的抉择和自己作出的抉择对总误差的影响这两个关键。

框架涉及整个研究项目设计的三个阶段：

1. 抽样前抉择。

2. 抽样抉择。

3. 抽样后抉择。

抽样前抉择是在进行研究设计时必须作出的抉择。这些抉择为抽样抉择打下基础。它们以一定的原理为依据，而这些原理也是分析总误差的依据。例如，有一种在抽样前必须作出的抉择是定义目标总体。目标总体的定义与研究目的有关。如果研究的目的是对社区为无家可归者提供的精神卫生服务设施进行评价，那么目标总体便可在逻辑上定义为于 1987 年 1 月到 1988 年 12 月间，在弗吉尼亚接受任何社区精神卫生服务的无家可归者。这一定义将在抽样抉择时使用，即在作出有关样本将从中选取的目标总体的清单编制的抉择时使用。目标总体和研究总体之间存在的差异将会引起非抽样偏倚。

抽样抉择是讨论抽样问题时通常都要讨论的问题。这些抉择中的每一个抉择将会直接和立即对总误差产生影响。任何一种抽样抉择，其样本选择的概率不是相等的，就是不等的。如果概率是不等的，且又没有用权进行修正，那么抽样偏倚便会因此而产生。

抽样后抉择需要就数据收集之后的一些程序作出抉择。其中一种需要我们作出的抉择是选择标准误差计算的方法。在通常情况下，标准误差的计算是以已经作出的抉择，特别是以已经选择的抽样方法类型为依据的。诸如这里提到的那样的抽样后抉择，将对由样本导致的总误差的大小的评估不无帮助，而另一些抽样后抉择，则会对总误差的降低有所帮助。

从这一实用抽样框架中的为数不多的抉择的简单介绍中，我们不难看出那些在整个研究过程中作出的抉择相互之间的关系。实用抽样设计法将与抽样有关的一系列抉择和研究设计及实施过程整合在一起，而不是将它们割裂开来。我们将与抽样有关的一系列抉择问题列在了表 3.1 中，并在本章下面的篇幅对这些问题一一予以解释。

表 3.1　样本设计中的有关问题

抽样前抉择

1. 研究的性质是什么——探索性的、描述性的还是分析性的?
2. 最感兴趣的变量是什么?
3. 研究的目标总体是什么?
4. 某些子总体或特定的群体是否对研究很重要?
5. 用什么样的方法收集数据?
6. 是否适宜进行抽样?

抽样抉择

1. 什么样的目标总体清单可作为抽样框使用?
2. 可容忍的误差或估计的效应的大小是多少?
3. 将采用什么类型的抽样技术?
4. 选择的概率是相等的还是不等的?
5. 选入的样本单位是多少?

抽样后抉择

1. 如何评估无回答问题的影响?
2. 是否需要加权?
3. 标准误差和与研究的估计值相关的置信区间有多大?

由这些问题表述的实用抽样设计所必须作的抉择与总误差直接相关。抽样前抉择主要与非抽样误差有关。不过有关是否对子总体需要进行分析的抉择则是一个例外。如果子总体分析是研究的目的之一,那么子总体分析的抽样变异性将受到样本中子总体的可能数目的影响。进行子总体分析所作的抽样前抉择,将会对以后的抽样抉择,如抽样方法的类型的抉择发生影响。而抽样方法则会对样本中的子总体的数目产生影响,进而对子总体分析的抽样变异性发生影响。

抽样抉择会对总误差的三个组成部分发生影响。例如,为目标总体编制的清单将会对非抽样偏倚的大小产生影响。选择的不等概的抉择可能会将偏倚引进样本。样本单位或对象的数目将会影响到抽样的变异性。抽样后抉择是我们估计、修正或分析总误差的各个组成部分的必经之途。

抽样前抉择

研究的性质究竟是什么——探索性的、描述性的还是分析性的

实施探索性研究的目的一般在于为正在进行的研究的课题提供方向或对研究本身有进一步的了解。它帮助研究者明确研究的重点，将注意力集中到重要的变量上，进而提出要检验的假设。描述性研究是许多抽样调查研究的主要目的。这样的研究的主要目的在于估计总体的特征、属性或态度。分析研究则要进行假设检验，考察群体之间和(或)变量之间的关系。实际上，许多研究试图兼顾描述和分析两个方面。对这一点能有所了解，并在遇有这样的情况时，确定二者孰轻孰重是很重要的。

在多数情况下，探索性研究都是一些准备性的工作，其目的在于将研究引向更为严格的描述性或分析性的研究。抽样方法通常都要在很大程度上受到资源和时间的限制。为了确保各种各样的群体都能被包含在研究中，我们经常会采用小的分层概率样本或立意配额样本。在许多探索性研究中，更为广泛的覆盖面比误差的降低更为重要，因为在探索性研究中，我们的目的并不在得到合理的，诸如平均数或比例这样的估计值。

描述性研究和分析性研究二者都牵涉到减少总误差问题。虽然就减少偏倚这一目标而言，二者并无二致；但二者对总误差中的标准误差部分的关注点却是不同的。描述性研究主要关注的是估计需要的精度。对需求进行的估价是政策取向研究的一种类型。这种类型的研究对估计的精度的关心是不言而喻的。例如，为了使项目能就需要得到服务和所需的特定服务项目类型的老人的数目作出抉择，样本估计值必须对即将递交的预算和服务设施计划是足够精确的，而这也是在这类研究中样本的唯一用途。而在分析研究中，我们的主要关注点则是确定对于我们期望的效应量，研究的力度是否足够到能拒绝零假设。达到这一目的的途径是强有力的检验(Lipsey，1989)。常用于评定某一项目是否对某一目标总体影响的评估研究也可被看做一类分析研究。

最感兴趣的变量是什么

选择变量和确定它们的变异性是在正式进行抽样设计前需要完成的一项重要工作。研究常常都会有多个目的。研究者可能预先想到文章中将会出现很多表格，或使用若干种检验假设的统计工具。最感兴趣变量的选择将会对在设计过程中稍后阶段的样本大小的确定有很大影响。

研究的目标总体是什么

研究的目标总体是研究者将要对其进行概括性陈述的群体。总体可以是一个个单独的人（北卡罗莱纳州的居民或洛杉矶的无家可归者）、个人的群体（里士满的住户或威斯康星的学校），或其他的单元（票据、州属的轿车或住宅单位）。将被检验的理论或正在研究过程中的政策将会帮助我们对目标总体做出定义。目标总体的定义也会经常涉及特定的时间段、地理位置、年龄和其他有关的标准。

某些子总体或特定的群体是否对研究很重要

研究者常常会对目标总体的某一部分予以特别的关注，对正在详察的某些现象做一些额外的分析。例如，某些正在考察收入维持实验（the income maintenance experiment，一种社会政策实验，译者注）的影响的学者可能对那些由单身的、有工作的女性担当户主的住户特别感兴趣（Skidmore，1983）。在考虑到子群体对研究的确很重要的时候，有若干种设计方案可供我们选择。而一个没有考虑到子总体问题的样本设计可能导致样本中的子总体成员过少，因而无法对之进行可信的分析。如果在抽样之前我们能够确定这些子群体，便可以用加大总样本量或不成比例地增加子群体的样本量来补救。

用什么样的方法收集数据

某些抽样抉择可能只能采用某种数据收集方法。例如，随机数字拨号法将生成一种随机选择的电话号码的样本。对这样的样本的调查只能通过电话进行。一个住户的概率样本主要用于实地调查，一般最好能用个别访谈法来收集数据。从行政记录中收集数据或用邮寄的问卷收集数据也要求某些特定的抽样设计。一般讲，邮寄问

卷总是会有比例较高的无回答的被调查人。如果样本成员中拒绝接受调查的人都有类似的特征，那么邮寄问卷中的高比例的无回答的被调查人将对抽样的变异性有较大的影响。

此外，数据收集工具的设计和使用达到的答案可信度也是抽样前应该考虑的问题。假定数据收集的工具是无偏的，我们要考虑的主要问题就是工具的可信性，工具的可信性越低，标准误差就越大。在确定样本大小的时候，工具的可信性是必须要考虑的问题。因为只有这样才可避免因为标准误差过大而影响统计结论的有效性。

是否宜于抽样

我们必须清楚我们为什么做出要进行抽样的决定。首先是因为，在一般情况下资源总是有限的。其次，在许多情况下，抽样调查得到的结果比总体普查得到的结果更精确。在多数情况下，在进行所有总体成员的调查时，许多资源都耗费在与总体的所有成员取得联系的过程中。初次联系的被调查人的回答率通常都不足50%，因而会引起严重的非抽样偏倚。从总体中抽样，初次联系需要的资源比较少，因而可以有更多的资源投入设计规定的各种后续的、用来提高回答率和降低非抽样偏倚的工作。此外，由于抽样方法的使用，减少了调查的个案，使我们可以有更多的精力投入数据处理，以提高数据的精确性。另一方面，在总体较小或调查信息将用于政治环境时，可能以不使用抽样为好。在这样的场合，我们在设计样本时，应该仔细考虑不使用抽样方法的原因，以使大多数异议能得到排除。例如，如果样本能提供地区性的估计值，那么在对项目进行评议时，可在一定程度上排除因样本未能包括每一个社区而引起的政治性异议。

抽样抉择

什么样的目标总体清单可作为抽样框使用

抽样框是将从中选择样本的清单。抽样框将为我们提供研究总体的定义，而目标总体和抽样框之间的差异则构成了非抽样偏倚。抽样框是总体，即研究者可以合理地进行推论的群体的操作定义。

一本电话簿可以作为一个社区总体研究的抽样框。它是服务区内所有已经安装了的正在运行的电话,并将号码刊登在电话簿上的住户的完整而清楚的抽样框。对于随机数码拨号法而言,抽样框是隐含的,而非直接显现的。这就是说,我们得到的是安装了电话的住户,而非总体的具体的清单。在很多时候,这种框架需要通过追问进一步提炼。追问的问题通常包括住处是否有目标总体的成员,或随机地选择住户中的成员,而不是默许选择那个接电话的人。

系统抽样、多级抽样和整群抽样同样也不需要编制整个目标总体的物质清单。系统抽样常常实施于具体的对象,如从一个文件柜中抽取单据,或从一个难民救济所的队列中选取个体。对于整群抽样和多级抽样而言,前者只需要编制完整的群的清单,而后者只需要每一级的抽样单位的清单。只有在多级抽样的最后一级,我们才需要列出目标总体的成员,而在此之前的那一级,只需列出要选择的抽样单位就可以了。

不完整的抽样框可能导致非抽样误差。最难以克服的不完整问题是抽样框中丢失了目标总体的某一部分。这样的情况一旦发生,将有可能产生某种无法用样本数据估计的偏倚。使用多种清单构建抽样框,选择一种不需要抽样框的方法,或对总体的缺失部分进行特别的补救性研究都是可供我们选择的、用来解决这一问题的办法。这些办法即使不一定能完全解决问题,至少也能使我们对它的严重程度有所估计。

可容忍的误差或估计的效应的大小是多少

在抉择的过程中开始考虑抽样的变异性问题时,我们必须首先确定可以容忍的误差或估计的效应的大小。对描述性统计而言,容忍的估计值的误差必须以我们最感兴趣的变量为依据。容忍的误差与围绕我们期望的带有一定的置信度的估计值的区间的大小有关。例如,一个从精神病医院出院的无家可归者人口百分比的估计值,可能必须在实际值的5%以内。对于某一临近选举的候选人来讲,在选前的民意测验中,可容忍的误差可能是1%。容忍的误差含有一个有关真值位于某一给定的区间内的可能性的假设。

在分析研究中,我们必须确定估计的效应大小。所谓估计的效应大小是指治疗变量(或自变量)可能给因变量造成的差异(此处原

文误将 estimated effect size 排成 estimated size effect，译者注）。我们来看一个司法方面的案例：法律要求对在犯罪过程中使用枪支的罪行追加一年的刑期。这样在犯有使用枪支罪的时候，我们可以预期的法律效应是在判决中增加了 12 个月的刑期。因而一个研究刑期实际增加的项目，其灵敏度必须达到能觉察刑期上出现的 12 个月的变化的程度。一个四岁儿童的发展项目效应，可以通过标准考试分上升或停留在初级的学生百分比的下降来评估，或者同时用这两个指标来评估。在这样的评估中使用的效应的大小，其变化程度，应该足以使项目的资助者判断他们对项目的资助是否物有所值。

可能的效应的大小和容忍的误差将要在计算有效样本量的时候使用。研究者的目的是在可以容忍的误差的范围内求得估计值，或进行能足够灵敏地察觉到估计的效应的分析检验。样本量是研究者达到这一目的的主要手段。但是抽样方法的效率将会对样本误差的大小和所需样本量的估计值有很大的影响。

采用什么类型的抽样技术

在第 2 章我们介绍了五种基本的概率抽样法：

- 简单随机抽样
- 系统抽样
- 分层抽样
- 整群抽样
- 多级抽样

虽然所有这些方法都是概率方法，但是它们都可能导致各自的抽样偏倚。选择什么样的抽样法取决于若干因素。这些因素包括精确的抽样框的可用性、预先得到的有关目标总体的信息的可用性、对于高效率的要求、对实施实地调查的要求，有时也会包括对电话调查的要求和目标总体的地理位置的要求等。不过抉择的过程并不因为抽样方法的选定而告终。

在抽样方法确定之后，我们还必须专门为选定的方法确定细节。如果我们选择了分层抽样，那么我们必须考虑我们需要用多少层？如果研究者选择了整群抽样法，我们还必须确定群可能有多少个。对于多级样本而言，研究者必须考虑怎样降低抽样的变异性问题：究

竟应当选择较多的含有较少的次级抽样单位的初级抽样单位呢，还是应该选择较少的含有较多的次级抽样单位的初级抽样单位？两种选择哪一种对减少抽样变异性更有利？在第4章我们将向读者介绍在实际工作中如何作出抉择的实例。在第6章我们将对各种备择方法的含义进行一番讨论。

选择的概率是相等的还是不等的

有关选择概率的抉择同样也会对抽样偏倚有所影响。例如对随机抽样而言，任何单独的单位的选择概率既可以等于抽样分数，也可以等于总体中被选作样本的那一部分所占的比例(n/N)。选择的任何单位的概率都是相等的。对于分层样本而言，任何单位的选择概率都是该单位所在那一层的抽样分数。使用分层抽样方法的概率既可以是相等的，也可以是不等的。多级抽样的概率的计算是最为复杂的。总选择概率是每一级的选择概率的乘积。每一级中的每一层和每一抽样单位都必须分别进行计算。

一个有相等的选择概率的样本叫做自加权样本。它表明我们不需要用权来对不等概问题进行修正。不等概选择是有偏的，因而在随后采用基于设计的方法(the design-based approach)来做估计和分析时必须要加权。所谓基于设计的方法是一种以设计的结构为根据进行加权估计或分析的方法。它可以对因设计而引起的不等概有所弥补。因为基于设计的估计和分析方法已为实际的抽样工作者所广泛使用，所以它将贯穿于本书随后的所有章节。对另一种备择的方法，即基于模型的方法(the model-based approach)感兴趣的读者，可参阅史密斯(Smith，1976)和卡尔登(Kalton，1986)的有关著作。

选入样本的单位是多少

研究者在样本大小问题上做出的选择，将会直接影响到抽样变异性。样本的大小取决于几个因素，在确定样本大小的过程中，从有效样本量的估计着手不失为一种明智之举。计算有效样本量是在某一特定的选择方法已经选定，且已在操作的意义上做出了具体规定之后，用于估计达到研究目的所需的样本量的一些方法。用于有效样本量计算的方法有两种，依据研究性质的不同，我们可采用其中的某一种。

对于描述性研究而言，问题的实质是在抽样方法已经给定的前提下，多大的样本才能得到有用的精度足够的估计值？这一问题与容忍的误差直接相关。我们可能还记得，容忍的误差只与抽样引起的变异性有关，而与总误差中的其他组成部分无关。估计值的标准误差是抽样变异性的量度。容忍的误差是估计值的标准误差乘以为希望的概率选择的 t 值的积，它是一个围绕着估计值的含有真值的区间。

$$te = ts_{\bar{x}}$$

式中，te 是容忍的误差；

t 是希望的概率的 t 值；

$s_{\bar{x}}$ 是估计值的标准误差。

容忍的误差是前面介绍的置信区间一侧的大小（$+ts_{\bar{x}}$）。因此在计算有效样本量的时候，容忍的误差实际上是研究者的研究可以容忍的置信区间的大小的一半。容忍的误差与估计值所需的精确度有关。

由前面介绍的计算置信区间的公式可知，置信区间的大小主要取决于三个变量：标准差、样本容量和统计值 t。此外，它也在一定程度上受到由有限总体修正（FPC）导致的抽样分数的影响。为了能从样本中得到对研究目的而言精确性足够的估计值，研究者可以直接控制的只有样本量，因为研究者能够对样本的大小进行调整。但是样本量的增加意味着数据收集的费用的增加。因此我们必须在精度和费用二者之间进行一番权衡，以找到较佳的结合点。而这正是我们必须要做的工作。

对于描述性研究而言，假定我们使用的是简单随机样本，那么样本量大小的计算就是标准误差计算的代数变换：

$$n' = \frac{s^2}{(te/t)^2}$$

$$n = \frac{n'}{1+f}$$

式中，n'是第一阶段计算的样本量；

s 是标准差的估计值；

te 是容忍的误差；

t 是希望的概率水平的 t 值；

n 是有限总体修正因子中的有效样本量；

f 是抽样分数。

上面讨论的容忍的误差将用于这一等式。容忍的误差是允许的标准误差乘以 t 值的积。换言之，它是为了构建置信区间而要从估计值加上和减去的值。因此这一等式中的容忍的误差必须除以 t 值，以使它表达为与标准误差等价的形式。在样本量决策中用到这一信息的例子将在第 4 章介绍（请特别关注表 4.6）。

考虑到这一公式的使用先于实际的数据收集这一点，我们便不难理解，公式中最难以得到的那部分数据是标准差的估计值。尽管如此，我们也并非完全是无计可施的，仍然还是有几种方法可供我们选择。这些方法包括预先进行一些研究、小规模的试调查和值域估计（estimates using the range）等。我们将在第 7 章对这些可供我们选择的方法进行讨论。

在主要以分析为目的的研究中我们用效力检验（power test）来计算有效样本量。效力检验主要用于确定某一特定的样本量是否在探测我们所期望的效应时是足够灵敏的（Lipsey, 1989）。拒绝项目或治疗是无效的零假设主要取决于效应的大小和估计值的标准误差的大小。标准误差越大或效应越小就越难以拒绝零假设。在零假设实际上是错误的，但我们却未能拒绝它；这种类型的误差叫做乙类误差（Type Ⅱ error）。

在争取得到有理由地否定零假设的机会时，样本大小再一次成为研究者手中主要工具。李普希（Lipsey, 1989）引进了一些其他的避免乙类误差的方法，这些方法从整个设计的灵敏度角度来考虑检验力。他在引进的同时对这些方法做了必要的说明。

如果在设计方案确定之后，样本量是影响估计值精度的主要因素时，那么在设计方案有所改动，尤其是在抽样方法有所改动时，我们可以用迭代法来考察这些改动对有效样本量的影响。在多级抽样时，改进分层方案或选择更多的初级抽样单位可以提高设计的效率。当然，这些改动也可能会使费用有所增加，但它的增加所需的费用可能会小于样本量增加所需的费用。

此外，我们也应该在这时考虑其他一些有关样本量的问题。例

如，设计使用的有效样本量得到的子总体成员数是否能满足对它们进行精度足够的描述的需要？在某一重要的子群体只是总体的一小部分的时候，我们可能无法将总样本量的数目增加到可以产生使该子群体有足够的成员数来满足分析的需要。这时，为了增加子群体的数目，我们也许可以采用不成比例的分层抽样，或补充样本这样一些设计。若采用这两种设计，我们不仅要考虑它们的费用问题，而且也要考虑它们的设计效率问题。

样本量的确定一般都是一个迭代的过程。该过程需要考虑和分析的因素有若干个，这些因素可能会改变我们此前作出的抉择。重要的问题是我们必须从总误差、因使用不同的抽样框而造成的研究总体定义的变化、满足研究目的的能力、时间和费用等方面对各种备择方法详加考察。为了确定一种备择方法的效应，我们应用框架中列出的每一项逐一对它进行考查。在对抽样方案作出抉择之前，最好也能对其他有关问题，如费用、调查员培训和对无回答的被调查人的追查等问题的影响加以认真的考虑。

抽样后抉择

如何评估无回答问题的影响

无回答就是未能从样本成员中得到有效的回答。在被调查人拒绝回答某一特定的问题，拒绝参与调查，或无法与被调查人取得联系的时候就会发生无回答问题。无回答是非抽样偏倚的一部分。无回答问题会导致研究总体和目标总体之间的差异。总体可以大致分为两个子总体：回答者的子总体和无回答者的子总体。无回答者子总体越小，它们可能对结果造成的偏倚就越小（Kalton，1983）。正因为如此，处理无回答问题的最好方法是消除无回答。福勒（Fowler，1984）介绍了几种减少无回答问题的方法。

因为完全消除无回答是不现实的，所以研究者能做的事是对它的影响做出评估。具体的做法是，就我们可以得到的那些总体值与样本特征和研究总体的特征进行比较。这样的比较可以有助于我们了解什么样类型的总体成员在样本中比例不够。然而问题在于，我

们可以得到的总体成员的特征，并不一定直接与那些对当前的研究很重要的变量有关。例如，像年龄、性别和种族这样的人口学变量常常是我们唯一可能得到的总体数据。此外，数据常常比较旧，因而可能不能反映当时的总体在这些特征上的分布。

不过，只要总体数据还比较及时，常可用来对样本数据加权，求得总体的比例，并考察加权对研究比较感兴趣的变量的影响。这种做法假定回答者和无回答者是没有差异的，充其量无非是对某些比例过低的群体所做的补偿而已。这一假定一般是未经证实的。

为了检验偏倚的大小，我们可以随机选取一个无回答者的子样本，对他们进行一些深入的追查以得到一些对研究很重要的变量的信息。有时我们把这样的追查叫做减员调查（attrition study）。在邮寄式调查中，我们可以通过电话进行追查（Dillman & Tarnai，1988）。从这样的调查中得到的数据可以与原先得到的数据做比较，以对由无回答引起的偏倚进行评估。我们将在第 8 章，展开对无回答问题评估的讨论。

样本数据是否需要加权

在研究者就抽样问题作出的抉择导致不等概，进而导致抽样偏倚产生的时候，为了进行必要的补救，一般都需要进行加权。有时某些分析需要加权，而另一些分析则不需要，是否需要加权，要视具体情况而定，不可一概而论。例如，一个研究可能会有两个分析单位，个人和住户。其中住户是等概地选择的，这样，当我们的调查对象是户的时候，就不需要进行加权。但是因为户内成员被选作调查对象的选择概率是以户内的成员数为依据的，所在以人作为调查对象时，便会发生不等概问题。于是必须要加权。我们将在第 8 章讨论加权问题。而有关加权问题的实例，将在第 4 章介绍。

研究变量的标准误差是什么

标准误差无论对描述性研究，还是对分析性研究都是很重要的。估计值的精确性和假设检验的灵敏度取决于标准误差。标准误差是抽样变异性的量度。标准误差的计算非常复杂，因为每一个统计量和每一种抽样方法的计算公式都是不同的。

估计标准误差的方法有两种：一种直接的方法，一种比较复杂的

需要用离差的近似值的方法。在实际研究中,直接的方法可用于简单随机样本、系统样本、分层样本和整群样本。直接方法的公式可用本章前面部分已经提到和讨论过的简单随机样本的均值的标准差的计算方法直观地加以说明:

$$s_{\bar{x}} = \frac{(1-f)^{1/2}s}{n^{1/2}}$$

这些公式是统计教科书中,在介绍各种假设使用的是简单随机样本的统计量时,阐述的主要内容。

其他抽样方法需要对使用直接方法的公式做一些修改。假定使用的是简单随机样本,那么实际的设计的抽样方差($s_{\bar{x}}^2$)与抽样方差之比则叫做设计效应(Kish, 1965)。这一比率反映了在其他条件相同的情况下,实际的样本设计在多大程度上增加或减少了总误差中的抽样变异性。分层可以减少抽样误差,因而它的抽样效应的值小于1。把一个较大的抽样分数分配给有最大的标准差的层,抽样误差和设计效应还会进一步降低。

整群抽样对设计效应的影响恰好相反,它会使设计效应的值大于1。之所以会出现这样的现象是因为,在整群抽样中,独立选择的数目是群的数目,而非最终选择的单位数。当:

> 群内重要的研究变量是异质(群内的标准差较大),或
> 群间的均值没有差别的时候,

设计效应就会减小。

因为群经常是在地理位置上定义的,所以为了使群内能有异质性更高的分组,我们必须使群变得更大。群的容量的增加势必导致数据收集的费用的增加。我们可以在选择前将群分层,这样做也可以使抽样的变异性有所降低。这就是说,在选择前,我们必须把群分到某一层,而用于定义层的变量必须适用于所有的层。

当然,我们也可以使用泰勒偏差逼近法(Taylor approximation of deviations method),或某种反复复制的方式来计算复杂的多级样本的抽样误差。但这些方法的使用只能依靠计算机的帮助。不仅如此,它们要求从每一层或每一个初级抽样单位中做至少两次选择。否则的话,我们必须做跨层的联合的选择。有关复杂样本的抽样误差的

计算，将在第6章介绍。

小　结

实用抽样设计的关键是权衡各种利弊作出正确的抉择。在不忘研究目的和资源有限的同时，尽可能地把总误差降到最小。在整个抽样过程中，抽样的设计和实施者固然应该随时作出适当的抉择。但是我们必须注意的是，为降低某一种误差而作的抉择，有可能会使另一种误差有所增加。

面对这复杂的来自多方面的挑战，研究者必须专注于总误差的减少问题。误差既可能由偏倚引起，也可能由抽样固有的随机波动引起。由偏倚引起的误差往往呈现某种系统性。我们无法将误差完全清除。比较切合实际的目标是尽力降低误差，周密的设计可以帮助我们达到这一目标。

本章阐述了总误差的三个组成部分：非抽样偏倚、抽样偏倚和抽样变异性。每一种都必须在设计过程中一一加以确认。本章也对那些在整个设计过程中必须解决的基本问题做了介绍。在这一过程中的每一阶段作的抉择都与其他的抉择彼此相关。下一章，我们将要介绍四个有关研究者在实际的社会研究中是如何作抉择的例子。

实用样本设计的四个实例

Four Practical Sample Designs

第4章

设计样本需要进行抉择，而不用实际的例子，可能是很难理解样本设计中的抉择问题及某一种设计的各种备择方法之间的关系的。离开了研究的具体背景，可能也是很难领会设计过程中对各方面情况进行权衡的确切含义的。在很多情形下，研究结果的使用者既无法从已经出版的报告中使已经废弃的备择方法复生，也不能据此完全理解设计的逻辑。因为这样的报告必定都是结果陈述取向的。

本书的主要目的是启发研究者在自己参与的样本设计过程中，对一系列可资利用的抉择及这些抉择在回答研究的问题中对样本有效性的意义进行思考。本章共列举了四个实例。这四个案例（其中一个系小型的非概率样本）对研究者遇到的各种情形、作出的抉择和某些被放弃的备择方法做了具体的介绍。

我们把介绍的重点放在了那些涉及美国州范围的研究。本人曾经参与过的大多数的研究都限于州的范围。这些研究的目的都是帮助州的立法和行政机构制定政策。从教育问题到高速公路的修建和政策的修正，州范围的研究和全美规模的调查一样，费用问题和减少总误差都很重要。州范围或州以下范围的研究的长处在于，它们不像全国范围的调查那么复杂，且有很多很有用的特点，便于我们从中进行归纳。不仅如此，它们还使我们在抽样框和数据收集方面有更多的选择，而这些问题对将实用抽样设计整合进整个研究过程都是十分重要的。

本章的最后一个例子是由密歇根大学的社会调查研究所的抽样调查研究中心（the Survey Research Center of the Institute for Social Research at the University of Michigan）设计的全国性的人口总体样本。

我们希望用这一例子加深大家对前面介绍的各种例子的理解。通过这一例子,我们希望大家能了解设计一个用于全国性个别访谈式调查的样本的复杂和艰难。与此同时,我们也试图用这一例子来说明在多目的和多年代调查中,如何在抽样设计过程中,在权衡各种利弊之后作出抉择。不过必须提醒大家的是,我们在这里介绍的这些内容,还不足以使你们能独立地设计一个面积概率样本(area probability sample)。读者在做这样的设计时,可能还需要求教于那些富有实践经验的抽样问题专家。

本书这一部分介绍的例子涉及的实际研究项目,在研究目的、数据收集的方法步骤和涉及的总体类型等各个方面都各不相同。第一个例子是北卡罗莱纳州居民调查。调查的目的是收集居民的意见和有关公共服务设施方面的信息。第二个例子是在佛罗里达州对年龄在75岁或以上的高龄老人进行的电话调查。第三个例子的样本设计则用于出院精神病患者的追踪调查。调查地点是弗吉尼亚,资料的来源是管理部门保存的记录。所有四个例子的概况可参见表4.1。

表4.1 样本设计实例

特性	北卡罗莱纳居民调查(1977)	佛罗里达老年调查	弗吉尼亚出院精神病患者追踪调查	全美抽样调查研究中心住户调查
目标总体	一般	特定	特定	一般
数据收集方法	电话,邮寄调查	电话调查	管理部门的记录	个别访谈
抽样框	报税单,公共医疗补助名单	随机数码拨号	出院病人清单	五个抽样框,每级一个
抽样方法	分层	二级分层	系统	多级
选择的概率	住户等概,个人不等概	地区不等概	不等概多级列举	整个住户(大致)等概
样本量	1 377	1 647	347	1 485
加权	户内的成年人数	地区人口数	出院人数	无人的户数

北卡罗莱纳州居民调查

1975 年,北卡罗莱纳州开始了一个对本州的公共事业、交通运输和经济发展情况进行考察的项目。为此研究者设计发展了一批量度指标。数据收集工作持续了若干年。整个调查涉及医疗卫生、就业和经济等多个方面。收集来的这些数据主要用于对该州已经实施的政策和项目的效果进行评定。

抽样前抉择

研究目的。调查旨在得到有关居民健康状况、就业和经济状况问题的可靠的估计值。得到的估计值必须是可作跨年度比较的。有鉴于此,样本产生的程序不仅必须是可以跨年度地重复的,而且也必须是足够灵敏的,以能捕捉到调查涉及的各个方面所发生的变化。不言而喻,这样的调查目的是要得到足够精确的描述性资料,来为政策的制定提供可靠的依据。“调查得到的数据将提供给州政府的有关部门使用。用途是多种多样的,包括制定规划、分配预算,制定政策和项目评估等”(Williams, 1982b, p. 1)。

数据收集的方法和研究总体。在这一例子中,有关数据收集法和研究总体的决策是密不可分的。调查结果将代表整个北卡罗莱纳州的成年人口,也就是说,该调查的目标总体是该州的一般人口总体。可供我们选择的能达到这一目的的方法有若干种:使用电话调查,用随机数码拨号抽样法。采用地区抽样法,给该州的每一地区分配一个选择的概率,然后对地区的全体或部分居民进行调查。也可以先收集编制一个北卡罗莱纳州人口的清单做抽样框。不管采用哪一种方法,我们都需要对研究总体和目标总体的定义之间的一致性程度做出评估。

第一种备择法,使用电话调查和随机数码拨号法是比较经济的。不过在北卡罗莱纳州这样的农业州进行电话调查可能会有比较严重的非抽样偏倚,因为在农村地区拥有电话的住户比较少。有关数字告诉我们,在调查开始的时候该州有电话住户的比例估计在 87% 左右(U.S. Bureau of the Census, 1975, cited in Grizzle, 1977, p. 3)。

不过，电话调查的费用比较低，且在调查的问题比较敏感的时候，产生的回答偏倚也低于个别访谈(Bradburn & Sudman, 1980)。若能辅之以更为深入的追踪调查，则回答率较低的问题也可大为改观(Fowler, 1984)。

面积概率抽样(area probability sampling)，即第二种备择法，是一种复杂的抽样方法，涉及一种地理单位逐渐变小的多级抽样。在面积概率的最后一级，首先是选择户，然后要从选出的每一户中选出一个被调查人。面积概率抽样通常都会加大抽样的变异性，因为生活在同一地理区域的人一般都有类似的特征。

为了使电话调查更经济和回答率更高，并尽可能地避免因排除了无电话的住户而引起的偏倚，比随机数码拨号法或面积概率抽样法更好的方法可能是设法编制一张北卡罗莱纳州居民的清单。然而问题在于，没有单独一种居民清单能全面到足以直接作为抽样框使用的程度。可以得到的清单，如记录1970年普查的住户资料的磁带上登载了住户的电话号码清单、住户与城市的用水管线连接的清单，以及城市的地名地址簿等都还没有全面到足以使非抽样偏倚降低到可以容忍的水平(Grizzle, 1977, p. 2)。不过研究人员发现，如果把含有户主信息的北卡罗莱纳所得税申报清单和1997年符合接受北卡罗莱纳医疗补助条件的居民名单合并在一起，估计便可将北卡罗莱纳州住户估计数的96%包含在抽样框内(Grizzle, 1977, p. 3)。而在1981年，该抽样框的覆盖率估计在94%左右。

抽样框清单给我们提供了研究总体的姓名和住址。为了降低费用，研究者设法得到了样本户的电话号码，并用电话调查作为主要的收集数据的方法。在选择的住户没有登录电话号码，或无法通过电话进行联系时，便代之以个别访谈。在所有完成的调查中，78%是用电话调查完成的，其余的则由个别访谈完成。

抽样抉择

研究总体定义为填写了1975年北卡罗莱纳州的所得税申报单的，或符合医疗补助条件，包括接受未成年子女补助住户中的个人。这两种清单涵盖了北卡罗莱纳的两大部分人，因为许多缺医少药的人都没有填写报税单。

采用的抽样方法是分层抽样，每一种清单就是单独的一层。在

初始样本中,在保持总体比例不变的前提下,从每一层各自选取了一个简单随机样本。因此住户的样本使用的是等概选择法,因而是自加权的。样本中的89%取自保税单,11%取自医疗补助单。在后一个的样本中,我们增加了样本中来自医疗保险单的比例,因为来自那一清单的被调查人无回答率高于来自报税单的被调查人。调整来自医疗补助单的比例使得这些户在总户数中所占的实际比例变成了13.5%(Williams, 1982b, p. 9)。

目标样本量初步设定为1 400个被调查人。考虑从每一户选择一个被调查人,并估计无回答和不合格(居住在另一个州,但在北卡罗莱纳填写申报单)率约为25%,我们实际需要从这些清单中抽取的户数为1 800到2 000。用于计算为达到要求的样本量而实际所需抽取的样本量的公式是:

$$n' = \frac{n}{1 - nr - i}$$

式中,n 是最终达到的目标样本量;

n'是从清单中实际抽取的样本量;

nr 是估计的无回答的比例;

i 是估计的清单中不合格的比例。

与这一样本设计和样本量相关的抽样变异性的近似值,可以十分方便地用计算比例的简单随机样本的公式计算:

$$s_p = \left[\frac{p(1-p)}{n}\right]^{1/2}$$

这个公式可用来计算抽样变异性的估计值。例如,假如样本的69%显示有某一特性,那么标准误差就等于:$s_p = (69 \times 31/1\ 377)^{1/2} = (2\ 139/1\ 377)^{1/2} = 1.25$。用于分层样本的,可以得到更为精确的估计值的公式,将在第6章介绍。不过对于只有两层的样本而言,这一近似值已经足够精确。表4.2显示的是1981年秋天的调查中若干变量的标准误差的估计值和95%的置信区间。这些数字都是在数据收集工作完成之后计算的。

表4.2中的标准误差显示,与比例关联的标准差越大,比例就越接近0.50。一个令医疗卫生政策制定者感兴趣的发现是,估计85%

北卡罗莱纳的居民将医生的诊室作为自己的主要医疗资源。研究者有95%的把握确信,以这样的方式使用医生的诊室的北卡罗莱纳居民的比例为83%到87%。

表4.2 北卡罗莱纳居民调查的标准误差和置信区间

变量	比例	标准误差	95%的区间 p 大于	 p 小于
$n=1\,377$				
户中有在职者	0.69	0.012 5	0.666	0.714
住户收入来自公共资源	0.48	0.013 5	0.454	0.506
工作满意情况	0.53	0.013 4	0.504	0.556
一年内可能失去工作	0.16	0.009 9	0.141	0.179
健康状况很好	0.86	0.009 3	0.842	0.878
医生诊室是主要的医疗资源	0.85	0.009 6	0.831	0.869
$n=459$				
在过去10年内移居北卡罗莱纳	0.42	0.023 0	0.375	0.465

那些移居北卡罗莱纳,并非世居的居民的子总体标准误差更大,因为这一特定子总体的数目较小(459)。移后北卡罗莱纳居民子总体的标准误差(2.3)几乎是其余比例的两倍。在列举的这一例子中,那些在过去10年中移居北卡罗莱纳的居民的比例估计为42%。该研究使用95%的置信度,所以过去10年中,移居到北卡罗莱纳的居民的子总体的比例,应该在总体的37%到47%。

抽样后抉择

在这一研究中,为了得到代表个人而非住户的答案,我们必须对答案进行加权。住户是等概地从清单中选取的,而作为被调查人的个人的选取的概率则取决于户内合格的被调查人数。例如,一人户中的被调查人被选中的概率四倍于户内有四个合格的被调查人。选择的概率是户内合格的被调查人的人数的倒数。权数通常是用下面的公式构建的:

$$w = e_i\left(\frac{n}{\sum e}\right)$$

式中，w 是权数；

e_i是户内合格的被调查人数；

n 是样本量。

例如，在总样本量为 1 377 和合格的被调查人总数为 4 022 时，我们将得到一个 0.342 的因子（$n/\sum e$）。这一因子乘以特定户内的合格人数便得到了修正抽样偏倚的权数。一人户的权数是 0.342，四人户的权数为 1.369。选择的概率是权数的倒数。因子 $n/\sum e$（在本例中是 0.342）的作用是使总样本量与权数之和保持相等，以便于显著性检验的计算。

加权的逻辑是以被调查人代表的人数为根据的。单人户代表户内的一个个人。选作四人户代表的个人代表了四个个人。因此四人户的被调查人需有四倍于一人户的权(1.369/0.342 = 4)。

我们必须认真地向读者报告来自调查数据的结果中可能存在的选择偏倚。表 4.3 列出了各类样本估计值与总体值的比较结果。当然前提是我们可以得到这些估计值的总体值。作者在报告中以这样的方式对样本的代表性问题进行了概括：“总的讲，1982 年秋天居民调查的样本，在主要的人口学特征的分布上……与全州的相应的估计值十分接近。”（Williams，1982b，p. 77）1982 年的比较结果比 1981 年的更接近。其原因至少有一部分应该归结于前面提到的对医疗补助清单的超比例抽样。

表 4.3　北卡罗莱纳居民调查被调查人人口学特征：1982 年秋天的数据和有关外部数据百分比比较

人口学特征	外部估计值	1982 年秋 NCC 调查	
		加权值	未加权值
年龄			
18～29	31.4	26.0	22.9
30～49	34.1	38.7	39.4
50～64	20.3	22.6	22.6
65 及以上	14.3	12.8	15.1

续表

人口学特征	外部估计值	1982 年秋 NCC 调查	
		加权值	未加权值
性别			
男	48.6	44.6	42.9
女	51.5	55.4	57.1
种族			
白人	78.1	76.5	78.8
非白人	21.9	23.5	21.2
家庭收入			
\$4 000 以下	10.7	5.9	7.6
\$4 001 ~ 8 000	11.5	10.5	12.1
\$8 001 ~ 12 000	15.9	13.6	14.3
\$12 001 ~ 20 000	21.9	30.4	29.9
\$20 000 以上	40.1	39.6	36.0
教育程度			
8 年和不到 8 年	18.3	15.4	15.9
上过几年高中	18.0	16.5	16.0
高中	35.4	40.1	39.0
上过几年大学	13.5	15.4	15.8
大学毕业	14.8	12.7	13.3
地区			
高山地区	14.7	15.8	15.3
山麓地带	53.7	54.0	55.0
滨海平原	21.5	20.4	19.7
海滨地带	10.2	9.7	10.0

来源：Williams，1982b。

佛罗里达高龄老人调查

定义为年龄在 75 岁或以上的高龄老人在佛罗里达,不仅总人数多,而且在总体中所占的比例还在不断上升。这些人最可能需要来自医疗补助体系的,包括长期住在疗养院这样的服务性支持。为了更好地对高龄老人提出的需求做出规划,并对有关住院治疗的备择方案进行考察,州有关部门的官员需要有关目标群体的信息。

抽样前抉择

研究的目的。1984 年,州的官员决定对高龄老人的需求进行评估,以对 1977 年和 1980 年做的评估进行更新。将需求评估用于对长期的医疗项目进行深入的研究在该州还是近年来才发生的事情。研究的主持人指出,其他那些研究"对接受各种康复治疗服务(HRS)的老人的情况——老人的健康问题、什么人在为老人提供帮助、老人的收入来源等做了深入的描述和分析。这些数据使我们看到了很多问题,如那些接受项目提供的服务的老人与那些未接受过项目提供的服务的老人相比,情况如何?大多数老人对诸如成人看护或家庭保姆这样的服务的需求是什么?"(Stutzman, 1985, p. 3)

该研究的主要负责人是玛丽·斯图茨曼(Mary Stutzman),而本例所列的资料的主要来源是题为《佛罗里达 75 岁及以上人口:基本数据汇编(1985)》的报告(*Florida's 75 + Population: A Baseline Data Sourcebook* (1985))。斯图茨曼(Stutzman, p. 16)将调查的两个主要目的概括为:

> 收集 75 岁及以上人口的人口学、健康状况和服务项目的数据。对 75 岁及以上的老人进行电话调查的可行性进行评估。

主要用于描述目的的变量经概念化之后可分为五类:

- 人口学特征
- 一般健康状况
- 功能能力和帮助
- 服务和社会支持
- 对进一步照料的要求

数据收集法。研究的目的之一是对电话调查是否能用于75岁及以上人口这一问题进行探讨。项目的工作人员认为老年人是不会接受电话调查的。以前的需求评估都是用个别访谈进行的。但是这次研究对费用问题考虑则比较多。作为一种备择的方法,个别访谈的交通费和调查费大大高于电话调查。以前的需求评估,因为受到经费限制,调查都是在非常有限的地区进行的,且延续了若干年。

在目标总体只占总人口中的一小部分(如本例中的6.5%),且没有可资利用的清单时,个别访谈的费用尤其高。在与1 000个住户进行联系之后,我们才能得到不到65户有高龄老人的住户。许多联系过的住户将被淘汰出样本,因为户内没有合格的个人,而这势必大大增加调查耗费的时间和交通费。虽然使用整群或多级样本进行个别访谈,费用则会有所下降,但是整群样本却会增加抽样的变异性。因此,为了使整群样本保持抽样变异性不变,我们需要进行数目更多的调查。苏德曼提出了一种用于在对稀缺人口进行筛选后,再进行个别访谈的方法。这种方法使整群样本设计的效率有所提高(Sudman, 1976)。然而,在电话调查是一种可供我们选择的调查方法,同时费用也是调查要考虑的重要因素时,对电话调查法做一番探讨无疑是很有价值的。

由于费用的问题,归结起来,实际上可供我们选择的收集数据的方法只有两种:使用客户清单的个别访谈和采用筛选性问题(Screening question)确定住户中是否有75岁及以上的老人的电话调查。选择第一种备择方法会直接对非抽样偏倚产生负面影响,因为未曾接受服务的高龄老人人口将会被排除在外。

抽样抉择

总体抉择。研究的目标总体是佛罗里达75岁及以上的人口。是否可以进行电话调查取决于我们能得到什么样的抽样框,而可供我们选择的得到抽样框的途径不外乎以下三个:得到一张75岁及以上的人口的清单、得到一张可以从中将75岁及以上的人口筛选出来的一般人口总体的清单、采用随机数码拨号法。

我们唯一可能得到的75岁及以上的老人的清单是那些正在接受HRS服务的老人的清单。如果使用这样的清单势必重蹈其他研究使非抽样偏倚加大的覆辙,因而无从回答上面提出的那些问题,显然我们不应该选择这样的抽样框。而第二种选择,使用一般人口总体的清单同样也是不可取的。因为那些有可能得到的全州范围的清

单都存在这样那样的选择偏倚。可作为抽样框使用的全国、全州，甚至地区性的精确的一般人口总体的清单一般都是无法得到的（Hess，1985）。例如电话簿往往都会遗漏那些未曾登录的电话号码、新近增加的电话号码和没有电话的住户。用电话簿作为抽样框将会导致人口中那些非常贫穷、非常富有和流动性比较高的人的比例过低。上面提到的北卡罗莱纳的例子不仅是一个例外，且具有一定的开创性，因为在许多其他州，提供类似的客户清单和用税收申报单编制清单是被明令禁止的。

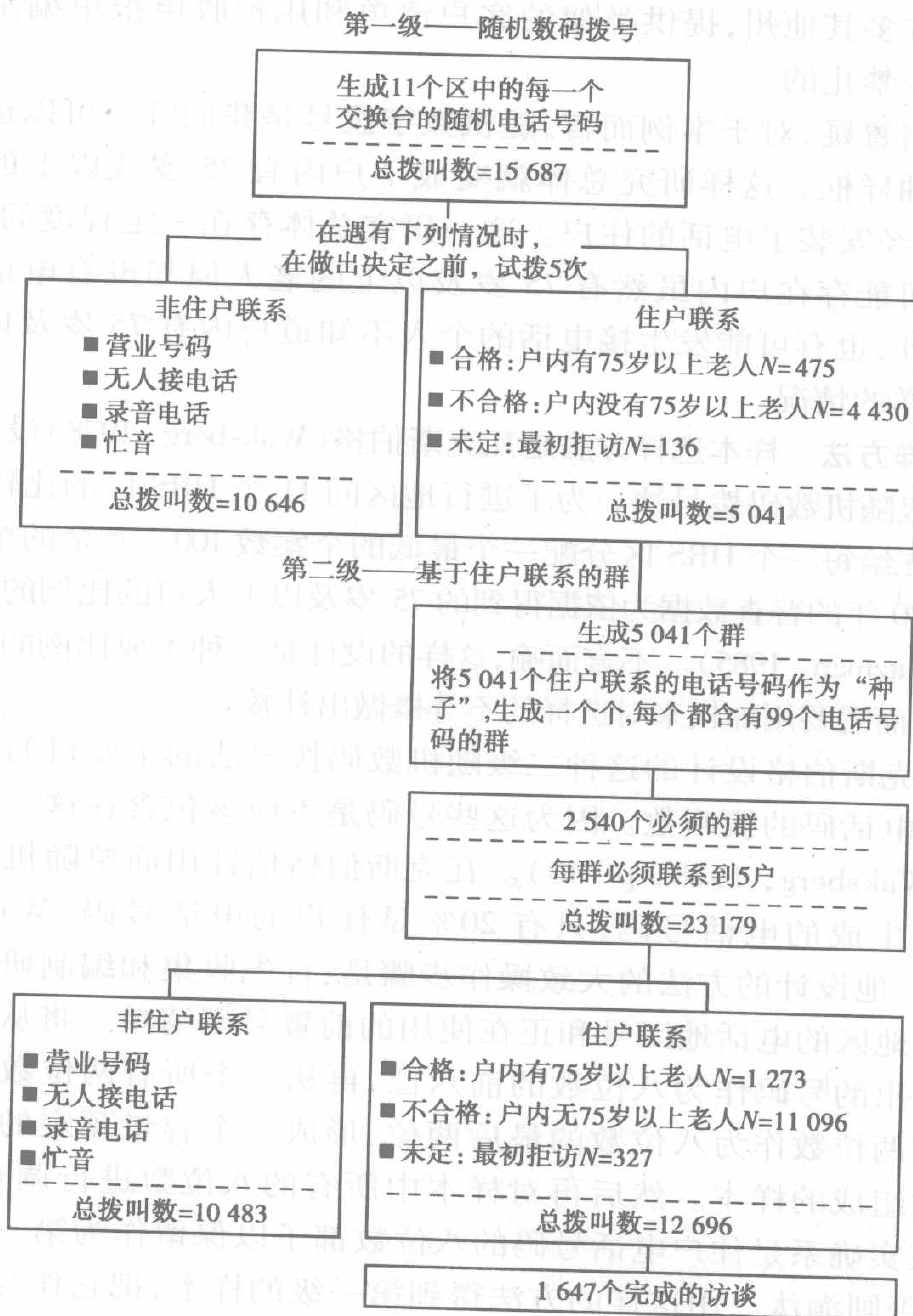
勿庸置疑，对于本例而言，随机数字拨号是我们唯一可以选择的可行的抽样框。这样研究总体就变成了户内有 75 岁或以上的老人的，且已经安装了电话的住户。这一研究总体存在一定程度的偏倚。因为有可能存在户内虽然有 75 岁及以上的老人但却没有电话的住户。此外，也有可能发生接电话的个人不知道户内有 75 岁及以上的老人这样的情况。

抽样方法。样本选择方法是瓦克斯伯格（Waksberg，1978）设计的分层的二级随机数码拨号法。为了进行地区间 11 个 HRS 区的比较，研究者决定先给每一个 HRS 区分配一个最低的个案数 100。其余的个案数，则以 1980 年的普查数据为依据得到的 75 岁及以上人口的比例的估计值分配（Stutzman，1985）。不言而喻，这样的设计是一种不成比例的分层抽样法，故而需要用加权来对选择的不等概做出补救。

瓦克斯伯格设计的这种二级随机数码拨号法的主要目的是减少非住户电话码的拨叫数。因为这些号码是不应该包含在这一项研究中的（Waksberg，1978，p. 40）。瓦克斯伯格估计用简单随机数码拨号程序生成的电话号码，只有 20% 是住户的电话号码（Waksberg，1978）。他设计的方法的大致操作步骤是：首先收集和编制研究者感兴趣的地区的电话地区号和正在使用的前置号码清单。将从清单中随机抽出的号码作为八位数的前六位，再从一个所有两位数的组合中抽出两位数作为八位数的最后两位，形成一个容量预定的由一个八位数组成的样本。然后再对样本中所有的八位数进行调查核实。凡经核实确系是住户电话号码的八位数都予以保留作为第一级抽样单位，否则淘汰。用这样的方法得到第一级的样本，把它作为一个母样本，因为在第二级抽样时，我们将通过随机取代每一个这样的八位数的后两位，生成一个个新的电话号码，并对每一个新生成的电话号码调查核实，凡核实后确系住户的号码都保留在样本中，否则淘汰。

最终得到容量为事先规定的数的样本(此段文字系译者为便于读者理解,重新编写,译者注)。上面只是这一方法的简单的介绍,凡是希望对这一方法的实际操作有更为详细了解的读者,请参见拉夫拉卡斯的有关著作(Lavrakas,1986)。

我们用这种方法得到了1 647个调查对象。而研究者确立的目标调查对象为1 500个。"为了使整个佛罗里达州的描述和估计的



资料来源:Stutzman, 1985。

图4.1 75岁及75岁以上老人调查:抽样策略

抽样误差保持在百分之三到百分之四（置信水平为 95%），我们需要的样本量在 1 500 左右”（Stutzman，1985，p. 24）。为了得到这 1 647 个被调查对象，我们对总共 38 866（15 687 + 23 179）个电话号码，打了 71 896 个电话。在所有这些打过的电话号码中，21 129（10 646 + 10 483）个是非住户电话号码，15 526（4 430 + 11 096）个是没有 75 岁以上的老人的住户号码（有关电话号码选择的更为详细的情况请参见图 4.1 和瓦克斯伯格法（waksberg method））。

抽样后抉择

我们需要使用两组权数来对这一产生不等概选择的方法进行补救。首先分层是不成比例的。表 4.4 列出了目标调查对象数、完成的调查对象数和分地区的加权个案数。第二组是对样本户中有一个以上合格的个人的被调查人选择法的补救所需要的权数。因为“聚类效应”与北卡罗莱纳居民调查相同，所以权数的计算公式也相同。

表 4.4 高龄老人调查：加权的个案

	高龄老人人口		实际样本数据			
区	人口[1]	百分数	完成的访谈数	样本百分比	在样本中的期望数	权数[2]
1	12 471	2.0	113	6.9	33	0.289
2	17 197	2.7	111	6.7	45	0.406
3	36 857	5.9	143	8.7	96	0.675
4	49 395	7.9	107	6.5	129	1.209
5	106 666	17.0	252	15.3	279	1.108
6	66 425	10.6	142	8.6	174	1.225
7	40 090	6.4	117	7.1	105	0.897
8	50 088	8.0	108	6.6	131	1.214
9	62 525	9.9	135	8.2	164	1.213
10	79 081	12.6	174	10.6	207	1.190
11	108 291	17.2	245	14.9	284	1.157
总计	629 086	100.2	1 647	100.1	1 647	

1. 资料来源：美国普查局。

2. 权数 = $\frac{\text{在样本中的期望数}}{\text{样本中实际的数}}$。

式中的样本期望值 =（总体%）×（总样本量），而总体% = 区内 75 岁的人数/州 75 岁总人数。

这时，我们也可以估计由无回答和其他原因导致的潜在的偏倚。在与独立的目标总体信息来源进行的比较无法确定选择偏倚不存在的时候，它可以为我们指明那些显然存在的问题（表4.5）。

表4.5 样本和总体特征：佛罗里达高龄老人比较

特 征	样 本	1985年官方普查项目
性别		
男	41.4%	40.0%
女	58.6%	60.0%
民族		
白人	86.7%	88.9%
黑人	7.0%	5.3%
西班牙裔	5.9%	5.4%
年龄		
75~79	46.7%	50.4%
80~84	30.6%	29.5%
85+	22.8%	20.1%

资料来源：1985年普查项目和1980年普查。一般人口特征：佛罗里达（PC80-1B4：Table 24）和Stutzman，1985。

在这一例子中，就我们收集到的总体特征值而言，没有发现明显的问题。样本中男性、白人和年龄在79岁以上人数的比例稍微高了一些。但设计允许的误差在3%到4%之间，所以我们可以认为样本与普查的比例之间虽然存在一定差异，但并没有大到需要进行额外的后分层加权的地步。

弗吉尼亚出院精神病人调查

20世纪80年代中叶，因为无家可归者，特别是贫民窟的居民人数的持续增长，国家把注意力集中到无家可归者和让精神病人从社会设置的看护机构中出院的政策的影响这两个问题上。判断让精神病人从社会设置的看护机构中出院这一政策的成败的关键，在于搞清社会设置的看护机构与社区提供的服务设施这两者之间的关系。1982年，弗吉尼亚立法机关需要在社会设置的看护机构的政策问题上得到技术

上的支援和有关这一问题的更新的信息。为满足这一需要进行的有关研究是由联合立法审计和审查委员会(Joint Legislative Audit and Review Commission)于 1979 年进行的一项研究的继续。立法者的目的在于在使用统计数据的同时,与来自各个方面的、比较熟悉这一问题的人士进行广泛深入的接触,倾听他们的意见,收集有关资料。

抽样前抉择

研究目的。这一研究的主要目的是对让精神病人从州立的精神病院出院的政策进行评估,并对社区提供的服务进行监控。

这项研究还有一个特殊的目的,那就是对有关机构在让自己的病人出院回到社区之前编制的计划的实施效果进行评价(Joint Legislative Audit and Review Commission, 1986)。编制诸如此类的计划的目的在于实现医院的服务与社区服务之间的转移。例如,由服务的提供者或负责出院病人治疗的医护人员参加的在出院前进行的协商会便在评估之列。

研究者希望考察病人的类型和出院后进入并在社区服务机构得到治疗的规程之间关系。从理论上讲,研究者的这一目的似乎使得目标总体的定义和数据收集方法的选择变得简单明了。目标总体是曾经在州立的精神病院住过院的已经出院的精神病人,而出院病人的数据则可以从社区服务提供者和医院保留的记录中收集。

实际上,要将总体定义操作化和使数据收集方法具有可操作性却并非一件容易的事。总体定义会因以下三个来自政策方面的原因而变得复杂起来:

- 精神病治疗类似一扇转门,因为一个病人可以进出医院多次。这样,一张出院清单上记录的内容是这一过程的“处理”点。病人有可能被处理若干次,因此有可能在清单上出现多次,他们的选择概率也因此而有所增加。
- 出院前的规程在 1984 年 7 月,即该项调查开展之前的几个月有了变动。我们无法用从执行老规程的记录中选出的个体对当前的采用的措施进行评估。
- 从社区得到的数据必须要有后续性。这些个体必须在社区已经逗留了四个月到一年,这样我们才能确定他们是否得到了

社区服务,进而再确定他们是否与社区的服务提供者保持着稳定的关系。

我们设计了一个包含了所有在 1984 年 9 月 1 日到 10 月 31 日间,从州立精神病医院出院的病人的抽样框,使得这三个问题迎刃而解。9 月 1 日这一个日子使得新的出院手续能有足够的执行时间。而 10 月 31 日这一个日子使我们得以得到:①足够的出院病人的样本;②病人出院之后至少有 4 个月的时间;③在选入样本的 350 个人中,只有 4 个人有多次出入院的记录。

数据收集方法。该研究采用的数据收集方法是从医院和社区保存记录中收集每一个病人的有关信息。为了比较全面地了解病人与公共精神病医疗系统之间的关系,我们收集了病人在 1983 年 2 月 1 日至 1985 年 2 月 1 日之间的历史情况。在数据收集中,我们特别关注的问题是样本涵盖的时间段内的出院记录,因而收集到的数据大多数都与出院和即将出院前的治疗情况有关。

数据收集涉及的记录历时两年,且散布在多个地点,因而需要一笔很大的费用。而保持记录的机密性又使本来已经够高的费用有所增加。在这样的情况下,时间和费用便会涉及对行政管理的记录进行深入评阅的需要,与此同时,也涉及为立法者提供技术支持的需要。在考虑样本量的时候必须兼顾这两个方面。

抽样抉择

在进行总体的定义和数据收集方法的选择的过程中,我们遇到了几个涉及抽样选择的问题。首先,研究总体的操作定义虽然已经确定,但是它对抽样偏倚的影响还有待进一步探讨。我们担心的主要问题是 9、10 月间出院的病人是否存在任何季节性的关联?而在对于一定数量的在那一期间出院的人进行考察之后发现,尽管可以证明存在一定的波动,但并不存在任何明显的季节性模式。不仅如此,独立的专家和实际工作者也未曾确定任何预想的季节性模式。因此我们认为出院是受系统中的其他因素,如医院的空间、社区服务的有效性等因素的影响的。而这些因素既不具有季节性特征,也不具有循环性特征。

其他的抽样抉择显然都受到环境因素的制约：

- 样本的大小是费用和时间的产物，而不是可容忍的误差的估计值或要求的检验力的产物。表 4.6 列出了总体和某一子总体的比例的期望的抽样误差的估计值。在期望的关系未能在分析的数据中发现时，我们应该考虑到这是否是因为检验力过低所致，对这一问题有所认识是很重要的。正是基于这样的认识，我们选择了一个容量为 350 的样本，它是切实可行的最大限度的样本数。表 4.6 显示了样本容量对置信区间大小的影响。最大的置信区间出现在比例为 0.5 的时候。表 4.6 也显示，随着比例逐渐趋向极端，置信区间也逐渐减小。

表 4.6 弗吉尼亚出院精神病人比例的 95% 的置信区间

整个总体的估计值			
比 例	样本量		
$p(1-p)$	300	400	500
0.10 – 0.90	±0.03	±0.03	±0.03
0.30 – 0.70	±0.05	±0.04	±0.04
0.50 – 0.50	±0.06	±0.05	±0.04
出院病人子总体			
出院病人比例	基于总样本量的子样本量		
$p(1-p)$	300	400	500
0.20 0.80	60	80	100
	(±0.13)	(±0.11)	(±0.10)
0.30 0.70	90	120	150
	(±0.10)	(±0.09)	(±0.08)
0.40 0.60	120	160	200
	(±0.09)	(±0.08)	(±0.07)
0.50 0.50	150	200	250
	(±0.08)	(±0.07)	(±0.06)

1. 每一格中子样本量和括弧中的置信区间都假定 0.5 的最坏情况。

置信区间：$1.96([p][1-p]/n)^{1/2}$。

资料来源：联合立法和审计委员会工作人员分析。

- 抽样框所含的信息使我们只能对病人出院的医院进行分层，而无法对病人进行分层。我们先对医院进行了排列，对它们做了隐性的分层，然后采用某种系统的方法选择样本。采用这样的方法得到的样本，是一个与医院出院病人数成比例的等概的出院病人样本。
- 在某种程度上讲，出院病人的等概样本是一个病人的不等概样本，因为某些病人在这一期间曾经有过多次的进出院的经历。实际上，在所有选择的病人中，只有 4 个有一次以上的出院经历，且基本上都是每人两次。

抽样后抉择

一般讲，无回答和由非抽样偏倚产生的问题是最难处理的抽样后问题。然而在本例中，无回答问题几乎不是一个问题。我们得到了样本中 350 个病人中的 347 个的数据。相对较小的样本量、合法地使用需要的记录和周密的追踪调查法几乎使所有的无回答问题都得到了比较完满的解决。这是该次调查的一个十分重要的特点。立法机关为这一调查下达的正式文件不仅使研究者能通过合法的途径得到有关资料，而且能得到实际部门的工作人员的非常好的配合。出院的病人是一个非常难以得到有关他们的追踪调查数据的群体。不言而喻，我们可以假定无回答群体将会显示与回答群体不同的某些特点。然而这并非我们对样本数据进行分析的要素。加权是另一个可能的抽样后问题。样本是一个出院病人的等概样本，但却是病人的一个不等概样本。对病人的极小的(4 个个案)不等概性进行加权后得到的结果与未加权的并无实质性的差别，故而我们在分析中删除了权数，分析的结果如表 4.7 所示。

表 4.7　弗吉尼亚出院精神病人

($n=347$)

	比例	s_p	95%的置信水平	
			大于	小于
人口特征				
男	0.58	0.03	0.53	0.63
白人	0.68	0.03	0.63	0.73
单身	0.82	0.02	0.78	0.86
自认失业	0.85	0.02	0.81	0.89
出院后精神状态				
要求治疗	0.78	0.02	0.74	0.82
要求监护起居状况	0.73	0.02	0.68	0.78
与社区服务站联系	0.63	0.03	0.58	0.68
与社区服务站联系至少 4 个月	0.40	0.03	0.35	0.45
子总体				
没有与社区服务站联系	0.37	0.03	0.32	0.42
	$n=127$			
搬家、接受私人服务，或返回监禁环境	0.39	0.04	0.31	0.47
拒绝服务	0.25	0.04	0.17	0.33
无联系	0.30	0.04	0.22	0.38

资料来源：Rog and Henry, 1986。

调查研究中心的美国本土住户样本

在 20 世纪 70 年代初，附属于密歇根大学社会调查研究所的调查研究中心(SRC)对他们自己采用的全美样本设计进行了修正。若干个采用个别访谈法的调查都使用这一样本。这一样本设计曾每年被用于 2～4 个研究，这些调查包括全美选举调查和消费者的财源、态度和行为调查(Hess, 1985, p. 19)。

抽样前抉择

研究目的。1970年修正全美样本的主要目的是提供一个可用于全美本土48个相连的州和哥伦比亚特区的住户或成年人口的概率选择的灵活的县的样本。在这样的情况中，样本并非只是为单独一次数据收集而设计的。选择的县及市、镇和农村地区将用于延续十年的研究。在此期间，在调查研究中心主要负责抽样工作的艾琳·赫斯(Irene Hess)女士承担了样本的修正工作。这里对这一实例的介绍主要以她的著作《社会调查的抽样问题，1947—1980》(*Sampling for Social Surveys*, 1947—1980)为根据。

样本设计的目的之一是为家庭研究提供等概的选择。等概的选择将不需要通过加权对选择的不等概进行补救，从而简化了数据分析的过程。对于调查研究中心的这一项目而言，这一性质尤其重要，调查的数据将向全美广大的研究者和教师开放，允许他们使用调查数据。选择的等概性将会极大地降低数据使用的难度。

样本设计需要为一个消费单位的户主的调查提供3 000个左右的调查对象。由于这样的设计采用了不成比例的选择，并需要从初始样本中筛选合格的被调查人，所以初始样本需要5 000到6 000个住户(Hess, 1985, p. 34)。设计需要的最大的单位数为5 000到6 000个的决定，主要是在考虑到这一样本设计的多个研究分析使用的统计程序类型和变量之后做出的。对原设计的改进首先是使精度有所降低，即在较低的精度就能满足研究的需要的前提下适当减少样本的数目。1970年进行的修改的另一个目的是尽可能地保持那些自20世纪40年代后期开始的，使用全美住户样本进行的一系列调查得到的数据的连续性。

数据收集法。而设计的另一个目的则是为采用个别访谈作为数据收集方法所规定的。出于实际的考虑，我们必须建立和保持一支调查员队伍，以能以比较经济的方式完成每年数次，每次为期六到八个星期的全国性的数据收集工作。全国性的个别访谈需要一支训练有素的调查员队伍，以能在有限的通勤时间内正确无误地完成数据收集工作和抽样程序。在研究设计要求对户内的居住者进行筛选，从中选出指定的居住者和进行跟踪调查时，通勤时间的限制问题尤

为重要。

研究总体。全美住户调查的样本的目标总体是美国成年人口。样本既可作为住户的样本,也可作为个人的样本。然而出于实际操作的考虑,设计对研究总体有所限制。首先,设计不包括阿拉斯加和夏威夷的居民。这两地的居民加在一起,不足美国人口的1%。保留总体中这么小的一部分人中的被调查人的困难程度,超过了将他们排除在调查之外所造成的影响。第二,总体中居住在军事基地的人口也被排除在调查之外,因为我们无法取得有关居住在特殊地点的人口数的资料。第三,诸如各种监狱、高校的宿舍和成人之家这样的集体居住单位也未曾包含在研究总体中(Hess, 1985, p.24)。最后,在最初的调查级段,经常需要对研究总体进行筛选,以使它能普遍地与目标总体相配。如全美选举研究就将那些并非美国公民和未达选举年龄的个体都筛选掉了。在筛选发现目标总体中存在一个以上的住户成员时,我们就必须使用一种客观的能使每个个体的选择都有一确定的选择的概率的客观的方法,而不要采用基于谁恰好可以接受调查就选择谁的基于偶然性的选择法,或基于调查员的主观判断的选择法。

抽样抉择

限于有关目标总体的注册或登记的信息的缺乏,全美个人和住户样本的抉择范围是十分有限的。不仅如此,不将总体相当大一部分的单位排除出去,就不可能编制可作为抽样框的清单,而这样巨大的工程所需要的费用是我们所难以承担的。因此,考虑到费用问题和面对面访谈的需要,为了利用概率样本所具有的各种优点,我们必须采用多级的面积样本。

唯一普遍用于多级样本的、可供我们选择的是一种综合随机数码拨号和电话调查这两种方法的调查方法。布拉德本和萨德曼(Bradburn and Sudman, 1980)曾对通过电话管理使用一种工具和通过个别访谈或邮寄式调查管理使用一种工具的相对效度问题进行过讨论。拉弗拉卡斯(Lavrakas , 1986)则对排除了总体的一部分的电话调查可能产生的偏倚和实施这一过程的实际可行性进行过讨论。

户的选择被设计为如表4.8所示的五级。在第一级,共选出了

74 个标准的大都市统计地区(Standard Metropolitan Statistical Areas, SMSAs)、标准的综合地区(Standard Consolidated Areas, SCAs),或县。其中的县都在标准的大都市综合区和标准的综合地区之外。第二级则在每一个初级抽样单位内各选择 3 ~10 个城市、城镇,或农村地区。第三级选择包括城市的街区、镇、"地块"或县中的小型地理单位的选择。然后选择群,最后则在选出的群中选择户。

有些研究在分析时需要将个体作为分析单位,为此我们特地增加了一级。在这一额外增加的级中,我们需要从户内合乎条件的人中选择一个被调查人。我们将在第 5 章介绍从户中选择被调查人的备择方法。在户内有一个以上的合乎参加调查的成员时,采用何种方法从户内选择被调查人将对调查的整个选择概率会有影响。这一情况与抽样框由作为研究的分析单位的个体的群或组的清单构成时的情况颇为相似。对选择的等概性的影响,在必要的时候是可以通过加权来补救的。

表 4.8　全美住户调查的样本的分级

级	单　位	简　介
1	县、SMSAs 和 SCAs[1]	2 700 个单位被置于 74 个层,10 个最大的 SMSAs 和 SCAs 是肯定入选的,一个初级单位选自 64 个其他的层。
2	城市、城镇和农村地区	从74 个初级单位中的每一个选取 3 ~10 个单位(平均 5 个),以大小分层。
3	城市和城镇的街区;农村的地块	从 370 个二级单位中最少选 3 个单位
4	期望的含 4 个住户的群	由总概率决定的选择数,导致群的选择的等概
5	住户	选择期望的 4 个住户中的全部或部分。保持选择的等概。
6	合格的个人	固定、客观的选择机制导致选择的不等概

[1] SMSAs——标准大都市统计区;SCAs——标准综合区。

诚如前一章所述,样本容量将会对抽样变异性和用样本数据计

算得来的估计值的精确性有很大影响。统计学家设计的计算方法使研究者得以在经费固定，或在确定的精确度的最小费用的前提下，使样本容量减到最少（Kish，1965；Sudman，1976）。在样本设计要满足多个调查的需要的时候，在确定样本容量时，我们必须考虑到三个实际的约束条件。

首先，因为使用同一抽样设计的各个研究感兴趣的题目有很大不同，所以各自感兴趣的变量和采用的分析方法也大不相同。为了使设计具有一定的弹性，以适应各种研究目的的需要，我们不一定要将样本量固定在能适合所有研究的水平上。其次，根据赫斯（Hess，1985）的论述可知，各种研究中每一种的最佳（即样本量最小）解决方案所要求的各种总体值（如总体标准差和调查费用的估计值）一般都是未知的，因为研究在很大程度上都是探索性的。最后，也许也是最重要的，也如赫斯所言，我们在对样本容量问题进行评估时，必须在抽样误差和非抽样误差之间进行权衡：

> 因为中心的许多住户调查不仅涉及的研究领域都比较新，而且涉及的时间较长，调查也比较深入，这样的调查常常都与非抽样误差有涉，所以样本容量比较小，调查对象人数通常在 1 000 到 3 000 左右，涉及的住户数与之相同。在新的和探索性的研究中，为了降低抽样误差而增加样本量通常都是得不偿失的，因为在这样的研究中，总误差常常取决于非抽样误差项（Hess，1985，p. 24）。

抽样后抉择

我们必须对来自研究设计的误差加以考察，以能确定它们在多大程度上对效度有影响。令人遗憾的是，估计非抽样误差的实际影响的工作在更大程度上是一项定性的工作，而非定量的工作。而对样本的抽样变异性所作的估计也不是那么精确，对于复杂样本来讲，情况尤其如此。

设计的非抽样偏倚可能是最大的误差源。在样本设计中可以观察到的非抽样误差有三种。第一种，在将样本数据与普查局最近发表的人口报告进行比较时发现，有一定数量的住户未曾包括在样本

中。虽然我们有意筛除了阿拉斯加和夏威夷的住户，以及那些生活在军事区域的人，但这不足以解释调查估计值和普查数据之间存在的4% ~9%的差异(Hess, 1985, p.240)。

第二种，住户中的人数存在一定数量的低报。在大多数情况下，只要有独立的估计值可以用做比较，通常我们都会发现全美住户调查中的年龄较小的类别的比例都过低(Hess, 1985, pp.246-257)。调查研究中心的调查员都被告知应该将那些居住在高校宿舍、军事基地和某些养老机构的成员从住户的常住人口清单中筛除。这一筛除加上被调查人的故意低报一起导致了户内人数的低报。

最后一种非抽样偏倚是无回答。尽管个别访谈的拒访率低于邮寄式调查，但是全美住户样本的拒访率仍然是相当高的。对于我们所选的调查研究中心在20世纪70年代使用的样本，拒访率平均为25.2%(Hess, 1985, p.59)。大都市地区的拒访率高于非大都市地区，而在大都市地区的中心城市的拒访率更高。

抽样变异性。全美住户样本的抽样变异性的总估计值是难以计算的，其原因不外乎以下两个:样本实际上在20世纪70年代的十年间抽取的若干个不同的样本，每一个样本都需要单独计算抽样变异性，尽管它们彼此相关。此外，抽样变异性是估计者和样本设计的函数。因此抽样变异性取决于正在研究的变量和使用的统计方法。

前面一章介绍的用于简单随机样本的计算公式，在用于更为复杂的设计，如像全美住户调查这样的调查设计时，未必都能产生精确的结果。面积抽样的使用增加了抽样的变异性，因为位于同一个区域的个体更易于具有共同的特征(Kish, 1965)。但是同质性的程度取决于在抽到的区域实际能发现的相似程度。那些因为面积抽样方法的使用而上升的标准误差可在一定程度上由分层而抵消。如我们在第6章将要介绍的那样的分层，主要依靠使选择更具异质性而使抽样的变异性有所下降。

在考察了简单样本和比较复杂的样本的抽样变异性之间的关系之后，斯图加特说道:“根据这些结果我们可以得到一条比较粗略的规则，那就是将非约束的随机抽样误差乘以1.25或1.50……它[粗略的法则]仍然具有一定的导向价值，如果调查的数据量相当大的话。”(Stuart, 1963, p.89)

复杂样本的抽样变异性的更为精确的计算方法已经设计出来了，它们对于更为复杂的统计量，如回归系数的抽样变异性的计算尤其有用。其中一种方法是一种平衡反复重复法。这种方法的关键在于得到反复运用样本设计得到的子样本的数据（Sudman，1976，p. 178）。将子样本的估计值合在一起计算总的样本估计值。通过筛除一个子样本，并确定余下的组合子样本的变异性来计算抽样变异性。不过反复选择的子样本将会降低可能用于设计的层的数目。在其他的条件相同的情况下，减少层就会降低设计的效率，抽样变异性就会因此而有所上升。

另一些估计抽样变异性的方法也已经设计出来了。这些方法在使用反复这一概念时，并未严格要求设计采用平衡的复制方法。其中一种称为半样本重复复制（the half sample repeated replication）的方法，将层中的观察配对，最大可能地保持原设计的结构。反复地从每一对观察中独立地抽取一个观察，选取一个个原样本的半样本。抽样变异性是整个样本的估计值和每一半样本的估计值之间差的和的平均数。

还有一种方法叫做折刀法。这种方法也与反复这一概念有关，但是每次只丢弃一个初级抽样单位。通过在一个层反复地丢弃一个初级抽样单位，对层中其他的单位加权，进而计算得到的统计值，可以估计该层对抽样变异性的贡献。为了估计总的变异性，我们可以把所有各层的变异性加总。不言而喻，使用需要反复计算的标准误差的估计方法是需要使用计算机程序的。

小　结

本章介绍的四个实例阐明了整个研究目的、数据收集方法、总体定义和样本设计的抉择这几个问题相互之间的关系。在这些例子中，这些抉择对于抽样后抉择和方法步骤的影响是显而易见的。这四种样本设计，从实用的角度阐明了不同情况下的样本设计问题。创造性的解决方案有助于对研究目的进一步理解。这些问题解决的具体方案经常是概率理论和诸如瓦克斯伯格法这样的抽样理论，以及抽样变异性估计法发展的结果。但是在抽样方法和实践的发展过

程中,对于研究发现的效度和减少不确定性问题的担忧,总是和在实际工作中对于数据收集的可操作问题和抽样设计中的费用问题的担忧连在一起的。正是这两种基本的担忧所产生的创造性的张力,产生了本书这一部分提出的每一种设计。

第5章 抽样框

Sampling Frames

抽样方案的抉择始于抽样框的抉择，通常需要在生成的若干个备择方案之间进行比较和评估之后才能确定。而这个过程肯定是非线性的。诚如本章所将要阐述的那样，有关抽样抉择的决策将会对执行早先的抉择的能力有影响。不仅如此，这些抉择也会对以后的备择方案的选择有所限制。这个过程是迭代的。

有关抽样框的决策会对源自实际的抽样设计的总误差的大小产生影响。目标总体和抽样框之间拟合的一致性将会减少非抽样偏倚，进而使总误差有所减少。某些抽样框抉择专属于使用的某些特定的抽样方法。正因为如此，虽然本书将抽样框的确定、抽样方法和样本容量这三个问题，分别在本章和下面的两章介绍，但实际上这三个问题是彼此相关的。

抽样框选择的评估标准包括以下几个方面：

- 总误差，包括非抽样偏倚和抽样变异性；
- 费用；
- 可行性；
- 对其他抉择的限制。

在讨论各种备择的抽样框之前，让我们先来区分一下目标总体的两种类型：一般目标总体和特殊目标总体。这一讨论可能会有助于大家对要讨论的问题的理解。一般总体通常由居住地和年龄来定义，例如加利福尼亚年龄在18岁以上的居民。上一章介绍的两个调查，如北卡罗莱纳居民调查和全美住户调查与全美罗伯民意调查(the National Roper Poll)和弗吉尼亚居民民意调查(Commonwealth

Poll in Virginia)一样,都属于一般总体调查。

与一般总体调查不同,特殊总体,如佛罗里达的老年调查和弗吉尼亚的出院精神病人调查的总体的定义都比较狭窄。通常它们都以个体或单位的有某些理论意义的具体条件或政策条文规定的目标总体定义的。

一般总体的抽样框

在通常情况下,研究者往往会发现,要寻找一张可资利用的清单,即使说不是不可能的,至少也是很困难的。在全国层次上,不存在可资利用的一般总体的清单。因为迁移、死亡和青年人成年等原因,一般总体始终处于不断的变动之中。而在地方层次,不管如何,可资利用的清单总还是存在的。一本最新的电话簿或一份水路或电路的用户清单常常是相当完整的。研究者无论使用这些清单中的哪一种,都将会丢失新居民和无家可归者。在考虑每种现成的清单中丢失的影响时,必须对照使用每种备择清单时的费用和完整性。

第一种可供我们选择的方法是编制一张有多种信息来源的清单。在北卡罗莱纳这一例子中,在编制一张更为全面的清单的时候,我们使用了两种清单。使用了报税单和符合医疗服务的人员名单这两种清单,研究者涵盖了该州约94%左右的住户。在地区性调查中,也许我们可以将私人财产或地方所得税申报单和社会服务的客户清单或符合医疗服务的客户清单合在一起使用。

在使用可资利用的清单组合之前,研究者必须对在去除了重复的条目之后,尚未包含的住户或人口的百分比做一个估计。此外,研究者也必须设法确定合并后的清单可能遗失的特定的群体这一点也是很重要的。在某些情况下,可能会有另外一张清单,他可以认定本来应该包含,但为以前的清单所丢失的群体的成员。在另外一些情况下,我们可以采用后分层——对样本量过低的组的成员进行后分层加权,使他们在比例上有代表性——也许可以在一定程度上化解这一问题。不管怎样,我们都应该对这一问题有所考虑,而对非抽样偏倚和数据使用带来的限制等问题的考虑则必须在清单合并法采用之前。

第二种可供我们选择的方法是使用一种不需要抽样框的方法。三种为我们所经常使用的不需要抽样框的方法是:随机数码拨号法、整群或多级抽样法和系统抽样法。

随机数码拨号法需要随机地产生一批随后用来进行电话调查的电话号码。瓦克斯伯格的两级电话号码生成法被用于佛罗里达的高龄老人调查。随机数码拨号法是我们得以避免在使用电话簿做抽样框时,因遗失未登录的号码和新登录的号码而产生的非抽样偏倚。但是,它还是未能包括那部分没有电话的高龄老人。不仅如此,随机数码拨号也只能应用于那些宜于使用电话调查的研究。

整群或多级抽样是第二种不需要抽样框的抽样方法。研究总体的成员的清单只是在最后一级的抽样时才需要,且只需要在紧挨最后一级的前一级的那些抽样单位中的总体成员的清单。例如一个四级的公立学校学生的样本可以设计成以校区为初级抽样单位,学校作为第二级抽样单位,教室作为第三级抽样单位,而把学生作为最后一级的抽样单位。在第一级,我们需要一张完整的校区的清单。在第二级,我们只需要那些在第一级实际选到的那些校区的学校的清单。我们只需要在紧挨当前这一级的前一级选到的那些单位的成员的清单,这就使得准备抽样框所需的财力、物力、人力和时间都大大降低。

在一般总体的研究中,由于很难得到比较全面的清单,因此比较常用的抽样方法是面积概率抽样法,它是多级抽样的一种形式。全美住户调查的样本就是面积概率抽样的一个很好的范例。在面积概率抽样中,每一级都需要有完整的清单。在选择初级抽样单位时,我们需要标准的大都市统计区(Standard Metropolitan Statistical Areas(SMSAs))、标准综合区(Standard Consolidated Areas (SCAs))和县(counties)的清单。而在二级抽样时,我们只需要那些在第一级抽样中实际选到的初级抽样单位中的城市、镇和农村地区的清单。采用个别访谈来收集数据,总是与面积概率抽样连在一起的。不仅如此,在样本容量相同的情况下,面积概率抽样的抽样误差一般是简单随机抽样的 1.5 到 2 倍(Stuart, 1963)。

系统抽样是最后一种不需要抽样框的抽样方法。不过,在不使用清单时,系统抽样要求抽样单位是实际在场的。不言而喻,我们很难找到切实可行的可用于一般总体研究的系统抽样法。但是对于特

殊总体来讲,这种抽样方法是很有用的。通常我们可以从市或区的客户办公室得到个案的文件,但是比较集中的清单可能不是现成的,或者不是定期编制的。此外,系统抽样对于其他未曾汇总成清单的发票或业务记录的抽样也是很有用的。

每种备择的抽样框都有自己的局限,在某些研究中可能不宜使用。如果数据必须通过个别访谈收集,那么随机数码拨号法可能就不是明智的选择,除非总体所处的地理区域的位置有限的。抽样框的选择或备择抽样框的选择必须考虑到与之有关的各种因素,如数据收集的方法、费用和总误差等。抽样框选择对非抽样偏倚和抽样变异性二者都会有影响。

特殊总体的抽样框

在三种得到一般总体的抽样框(采用一张现成的清单、合并两张或更多张的清单和使用一种可避免使用清单的方法)方法之外,我们还可以增加第四种备择的方法,那就是编制一张清单。对于那些马上就可以确认的总体,我们有可能编制一份清单。例如,我们可以要求全州的每一校区提供有关的信息,编制一张全州在校上课的学生的清单。我们也可以用一个博物馆保存的参观者的日志编制一张博物馆参观者的清单。

研究者在设计特殊总体的样本时,应该尽可能避免未经思考地使用一张现成的清单。现成的清单虽然为我们提供了一张便于使用的抽样框,但是它们可能会将相当大一部分总体成员排除在外。在很多情况下,这样的情况都会引起非抽样偏倚。在老年人需求评估中,使用现成的接受公共提供的医疗和社会服务的客户清单,将会排除那些不接受公共服务和接受私人服务的老人。

职业组织和协会一般都保留自己的成员的清单。它常常会诱使研究者把它们作为抽样框使用。如果这些成员全部都在目标总体之中,或目标总体不含任何非组织成员的个人,那么这的确是一种不错的选择。例如,将一张全州教师组织成员的清单用于该组织成员的研究的确不失为一种明智之举。然而,假如目标总体是全州的教师,那些可能与教师组织的成员有着显著差别的非教师组织成员的教师

就可能被研究所丢失。

如果研究希望将一张现成的清单作为抽样框，那么就必须对它是否会引起非抽样偏倚这一问题进行仔细的评估。我们必须结合可行性、费用、对其他抉择的限制和总误差等问题，对备择的抽样框进行探讨和评估。可行性问题是一个颇为特殊的问题，它牵涉备择的抽样框是否切实可行？例如，拉夫拉卡斯（Lavrakas, 1986）认为在目标总体不足一般总体的10%~20%时，随机数码拨号法便是不可取的。佛罗里达老年需求评估的目标总体（是一般总体的6.5%）可以说是用随机数码拨号法进行的占一般总体百分比最小的电话调查。为了完成1 647个调查，调查需要拨打将近72 000个电话。尽管打了这么多的电话，但是完成每一个调查所需的费用仍然比较低，只有40美元，大大低于个别访谈所需的费用（Stutzman, 1985）。此外，如果采用了随机数码拨号法，那么我们使用的收集数据的方法只能是电话调查法，其他各种实用调查方法都无法使用。对需求评估而言，把电话调查用于这一目标总体进行测试也是它的目的之一。

总误差和抽样框

抽样框可能会对抽样设计的总误差有所影响。尤其应该引起我们注意的是，抽样框是引起非抽样偏倚或目标总体和研究总体之间的差异的两个主要原因之一。在使用抽样框之前，研究者应该对抽样框中存在的四个潜在的缺陷可能对总误差带来的影响有所了解：

- 丢失：抽样框中丢失了目标总体的某些人口单位。
- 重复：某些单位在清单中登录一次以上。
- 不合格：抽样框中有不属于目标总体的单位。
- 整群登录：抽样框中的单位以群体形式登录。

这四种缺陷的每一种都会引起非抽样偏倚，从而使总误差加大。这些缺陷也同样存在于那些不需要实际的目标总体清单的方法之中。此外，那些不需要实际清单的备择方法，可能会因为使用的抽样方法（整群或多级抽样）加大了标准误差，而导致总误差的增加。

丢失。所谓丢失是指目标总体的成员未曾被包含在抽样框中的情况。全美住户调查丢失了阿拉斯加和夏威夷的居民，以及生活在公共机构中的居民。这些丢失属于一种已知的丢失。它使来自样本的总户数的估计值低于美国普查局的估计值5%～6%（Hess, 1985, p.58）。不过，这些差别并非都由已知的丢失所致。

丢失是目标总体和调查总体之间的差异，因而它是非抽样偏倚的起因之一。丢失就其实质而言，它是样本框中最难以弥补的缺陷。已知的丢失大多都可以得到纠正，如果它们所引起的偏倚的确是值得纠正的话。阿拉斯加和夏威夷的居民是可以包括在全美住户调查中的，他们之所以未曾包括在调查中，主要是因为为这么小的一部分全美的一般总体，而专门聘用和供养一批训练有素的调查员费用过高。

那些无法确认或无法联系的丢失，不论增加多少费用也难以将他们去除。北卡罗莱纳居民调查的抽样框丢失了6%左右的总体。然后究竟是什么样的居民被丢失了却不是十分清楚。因而我们也不知道怎么样来改进我们的抽样框。在遇有这样的情况时，抽样框必须对照其他的备择方法，如随机数码拨号和面积概率抽样进行评估。

评估抽样框的四个标准是：总误差、费用、可行性和它对其他研究抉择造成的限制。实用抽样设计就是产生于对相关的备择方法进行比较评估之中的，而非产生于对抽样框进行的纯粹的理论估价之中的。然而，对各种备择的抽样框做出的选择，最终必须要能将总误差限制在可以容忍的范围之内，否则研究的结果可能无法使用。为克服减少误差和增加费用之间存在的矛盾而做出的种种努力会促使样本设计的不断发展，而前面引用的"洛杉矶无家可归者研究"便是一个很好的例子（Burnam & Koegel, 1988）。他们采用的策略是先列举所有的无家可归的人，然后将他们分成三层，再从各个层中抽取样本。这三个层是在临时收留所过夜的人、在收容所进餐但没有床位的人和被收容在室内过夜超过一个月以上的人。

重复。每当总体的一个成员出现在样本框中一次以上时，重复问题就发生了。重复问题的例子可以说是不胜枚举的。在随机数码拨号中，有一部以上电话的住户都会有重复。在将有多次登录的住户的电话簿用作抽样框时，也会发生重复问题。在将两张内容有一

定重叠的清单用于北卡罗莱纳居民调查时同样也会发生重复问题。

在抽样框记录的是一项项业务，而非一个个个体的清单时，最后一种的重复问题就会发生。例如，让病人出院这一研究使用从公共机构出院记录清单作为抽样框。出院是办理的一种手续，某一单独的个人有可能办理几次这样的手续。一个单独的个人办理的多次手续便是重复。由于调查涉及的时段很短（两个月），所以只有数量很少的重复进入了这一研究的抽样框。

重复问题既可以在抽样之前处理，也可以在抽样之后处理。处理的方法是将重复的条目从抽样框中去除。但是对于大的总体而言，这一程序可能是相当费钱和费时的。

我们也可以在样本选择之后对重复问题进行补救。重复问题可以看做因为重复登录而增加的选择的概率。下面便是一个重复登录的简单例子：

Ike
Mary
Don
Miguel
Debbie
Assad
Juanita
Ike
Jimmy
LaFarn

这份 10 个姓名的清单中包含一个重复的条目，Ike 被登录了两次。假定从这一张清单中选择了三个学生。Juanita 有概率 0.3（3/10）或有 30% 的可能被选入样本。Ike 因为被重复登录，致使他的选择概率大于 Juanita。实际上，Ike 被选择的概率是 0.53。

选择的不等概造成的抽样偏倚可以用加权来补救。在这一例子中，权数应该是名字出现在清单上的次数的倒数：1/2 或 0.5。这一权数对许多个体因被登录了两次而有两倍出现在总样本中的可能性这一问题作了补救，以使样本中的比例能与总体保持一致。权数使用不当将会产生严重的后果。如我们常会遇到含有屡教不改的多次

出入监狱惯犯的出狱犯人清单,或含有多次住院出院精神病人的精神病人的清单。如果我们用病人出院记录清单作为抽样框,而又未对多次出院的病人的选择概率进行修正,那么那些有多次出院记录的病人的比例就会过高。

为了确定权数,我们必须确定样本中每一成员的登录次数。而这样的工作,对于大的抽样框和大样本来讲,工作量是十分巨大的。在抽样框是以唯一的标识码自动登录,且重复条目又可以用计算机进行识别的时候,工作量就会大大减轻。对于随机数码拨号法而言,我们必须在调查中加上一个类似户中究竟连接了多少电话线路这样的问题,然后用这一数目作为对答案进行加权的分母。

不合格。不合格是指那些虽然出现在清单中,但并非目标总体成员的单位。北卡罗莱纳研究的不合格可能包括那些在清单编制完成、调查开始之前已经搬出该州的个体。在高龄老人研究中,不合格是指那些虽然已经联系到但户中却没有 75 岁以上老人的住户。

只要有可能,我们就应该设法将出现在抽样框中的不合格数降低到最少的程度。从不合格的调查对象中收集数据将会耗费本来应该用于从合格的被调查人收集数据的资源。不言而喻,这样的耗费,个别访谈将大大高于电话调查。在邮寄式调查中,不合格问题则更为严重,因为为了达到要求的回答率,我们需要更多的费用来进行追踪调查。不仅如此,为了更为精确地计算无回答率,而将不合格的被调查人与合格的无回答的被调查人加以区别几乎是不可能的。

在无法将不合格者从清单中去除的时候,他们必须在不去除任何合格的被调查人的前提下被有效地筛去。表 5.1 显示了在佛罗里达高龄老人调查中使用的筛选方法。这种类型的筛选法使研究者得以将不合格和拒访加以区别,进而对因拒访而引起的可能的偏倚有所估计。在个别访谈和电话调查中,在一个正式的调查开始之前,我们常常会对个人或住户通过双重测试进行筛选。首先,被调查人会被问一个或一串问题,以便调查员据此对被调查人的合格性做出判断。如果调查员断定,被调查人是不合格的,在多数情况下,该被调查人就会被直接问及,户内是否有任何属于目标总体的成员。不言而喻,在邮寄式调查中,我们只能依靠被调查人自行对合格性问题做出的判断。

清单中不合格条目的存在并不会导致结果的偏倚，除非合格者被不经意地筛除了，或将不合格者误作合格保留在了样本中。不过在设计过程中我们必须对不合格问题有所估计，并如像第 6 章所示的那样，对从抽样框中选择的单位数做出修正。例如，假如样本需要的个案数为 1 500，而样本框中的含有 20% 的不合格单位，那么抽取 1 500 个单位，才有望得到 1 200[1 500 ×(1 - 0.2)]个合格的被调查人。为了弥补这一问题，我们必须从清单中抽取 1 875[1 500 ÷(1 - 0.2)]个单位。

表 5.1 筛选高龄老人

开始时间：________

您好，我是________，我在塔拉哈西(Tallahassee)的佛罗里达大学给您打电话。我们的大学目前正在进行一项有关 75 岁及以上的居民的服务需求的调查。

首先，我想确认一下，我拨叫的电话号码是不是……？

因为这个电话号码是由计算机随机选择的，所以我需要了解一下，这个号码是住户号码，还是办公号码？

办公号码——谢谢，给您添麻烦了，我们的调查对象只是本州的居民，不多打扰了。

住宅号码——下一个问题。

您家中是否有 75 岁或 75 岁以上的老人？

没有——谢谢，给您添麻烦了，我们的调查对象都是 75 岁及以上的老人，不多打扰了。

有——下一个问题

整群登录。许多抽样框是分析单位成组的清单，而非单位自身的清单。一般总体调查常将住户而不是个人作为抽样单位。在第 4 章介绍的三个例子中，情况就是这样。这三个例子是居民调查、高龄老人研究和全美住户调查。住户是某些变量单位而个人则是另一些变量的分析单位时，这种现象尤其容易混淆。所有这三个例子都是用了不同的分析单位，户是一些变量的分析单位，而个人则是另一些变量的分析单位。

问题不限于以户为基础的抽样框。在政策研究中，个案常常是

列在抽样框中的抽样单位。而一个个案既可以是由一个个体组成的,也可以是由几个个体组成的。例如,一个社会服务机构可能保留了登录寄养(foster care)个案的清单,而个案则可能定义为居住在提供寄养的住户中的人。住户中接受的寄养儿童数,以及成年和非寄养的儿童数在各个个案间可能有着很大的差异(儿童寄养是美国实施的一项儿童福利项目,译者注)。

如果清单的条目是以群编制的,那么研究者需要解决的问题有两个:①选取一个或一些被调查人,②确定选择的概率。在单个条目中包含一个以上的合格的被调查人时,我们可以对所有合格的调查人进行调查,也可以选择其中的一个进行调查。究竟哪一种方法更有用则取决于研究的目的。如果选择了所有合格的被调查人,那么样本应该被认为是一个整群样本。

被调查人的选择可以通过衡量事先规定的某种标准来进行,如户主,或前例中接受寄养儿童家庭中的最年长者等。不过在将这些方法运用于个人时,有可能使结果产生偏倚。用户主作为标准来选择被调查人,可能会导致中年男性超比例。采用谁来开门或接电话谁就是被调查人作为标准来选择被调查人,同样也会造成偏倚:“调查员的判断力、被调查人的判断力和可遇性(它与工作状况、生活方式、年龄有关)将对什么人会成为被调查人有影响。”(Fowler, 1984, p. 33)

一种最初由凯思(Kish, 1965)设计的、比较可取的方法是将选择的过程随机化。这一方法先将合格者列表,并依据一种统一的标准给他们编号,然后从事先提供的专门用于选择被调查人的随机数码表中选择一个数,列表中与该数对应的那个人就是选中的被调查人。这种方法仍不适用于邮寄式调查。

有两种方法可作为凯思法的备择方法。在将凯思法用于电话调查时,选择过程需要的时间过多,从而导致被调查人拒绝接受调查,进而引起非抽样偏倚。特劳尔达赫-卡特选择法则可以缩短选择的过程(Troldahl-Carter, 1964)。为了选择一个一般总体的样本的被调查人,需要问的问题只有两个:

- 包括你本人,您家中共有多少个年龄在 18 岁及以上的人?
- 他们中间有多少个是男人?

根据被调查人对这两个问题给出的答案，调查员根据被调查人可能给出的答案构成的矩阵来选择一个被调查人。为了使选择保持平衡，用于一般总体研究的矩阵共有四种版本。表 5.2 便是这种矩阵的一个范例。在选择的结果是最年长或最年轻的人时，不论被选择者的性别是什么，这种方法都稍微会产生一些偏倚。不过这个问题只有在户内有两个以上相同性别的成年人时才会出现。同样，这个方法也会导致样本中女性超比例。在一项研究中，不同的选择方法或提问方式的拒访的差异是微乎其微的：凯思法为 7.8%，特劳尔达赫-卡特法为 7.2%（Czaja, Blair, & Sebestile, 1982）。

表 5.2 特劳尔达赫-卡特选择

男性人数	户内成人数			
	1	2	3	4 或更多
0	女	最年长的女人	最年轻的女人	最年轻的女人
1	男	男人	男人	最年长的女人
2	—	最年长的男人	最年轻的男人	最年轻的男人
3	—	—	最年轻的男人	最年长的男人
4 或更多	—	—	—	最年长的男人

第二种方法通常被叫做“最近生日”选择法。这种方法用一个问题来确定被调查人：“现在住在您家中年龄在 18 岁或以上的人当中，谁最近要过生日？”奥洛克和布莱尔（O’ Rourke & Blair, 1983）发现，在使用最近生日法的时候，只有 1.8% 的被调查人在选择过程中终止了调查，而在使用凯思法的时候，这一数字为 4.1%。这一方法也是系统的，而非随机的，同样也会在实验过程中产生可观察到的偏倚。

从一条只含一个合格者的群的条目选择一个被调查人的选择概率，与从一条含有三个合格者的群的条目选择一个被调查人的选择概率是不同的。第一种情况的选择概率是第二种的三倍。我们可以通过加权来去除因选择的不等概而引起的偏倚。前一章提到的北卡罗莱纳的研究为我们提供了一个类似的加权设计的实例：分配的权数是选择概率的倒数。权数 1 分配给了只有一个合格者的个案，权数 3 则分派给了有三个合格者的个案。我们用这个实例介绍了更为

复杂的加权方法，我们对加权的样本容量做了修正，以使它与未加权的样本容量均等。

结　论

在作出了抽样前抉择之后，抽样框的决策是抽样设计的第一步。就抽样框而言，有三种方法可供我们选择：①使用一张现成的清单；②将若干张清单并在一起或编制一张清单；③使用一种不需要清单的方法。对抽样框所作的抉择规定了研究结果可以推论的总体。在研究总体和目标总体同义的时候，我们便去除了非抽样偏倚的一个可能的来源。而从实际情况看，这样一种高度的同义几乎是不可能的。因此，抽样框的决策是在权衡各方面的因素之后做出的。我们无法在改进抽样框的同时减少抽样的变异性，如增加样本量。二者都对总误差有影响，我们究竟应该在哪一方面花费更多的资源，必须权衡利弊，小心行事。

不仅如此，有关抽样框的决策常常会对其他的研究抉择有所影响。例如，随机数码拨号实际上使电话调查成为研究者唯一可采用的调查方法。抽样框的抉择必须要与在设计过程中作出的其他抉择联系在一起考虑。我们应该对各种可供选择的方法加以比较。比较的内容包括：得到抽样框所需费用的高低，使用该抽样框收集数据所需的费用，得到和使用该抽样框的可行性、总误差（偏倚和变异性），以及采用该抽样框对其他的研究抉择的影响等。

抽样框的选择常常是对一个又一个的备择抽样框进行非正式的考察之后做出的。有关抽样框的深思熟虑的决策是在其他的设计抉择，如抽样方法和样本容量的假设之上作出的。例如，在抽样框的决策涉及抽样方法的时候，我们必须先就此做出一个假设，然后对它进行分析。如果假设未能被证实是比较正确的，那么我们就应该对抽样框进行修正，然后再重新开始假设检验的过程。

第6章 抽样方法

Sampling Techniques

在抽取概率样本时,可供研究人员选择的抽样方法有五种:

- 简单随机抽样(simple random sampling)
- 系统抽样(systematic sampling)
- 分层抽样(stratified sampling)
- 整群抽样(cluster sampling)
- 多级抽样(multistage sampling)

不过在实际抽样过程中,为了能做到真正切实可行,每一种方法都需要反复斟酌,作出很多抉择。为了降低总误差,我们常常需要将这几种方法混合使用,而这就会使问题变得更加复杂。例如,如果群是可以以某一个重要的变量有序地排列的,那么我们就可以将系统抽样用于整群抽样,从而使我们能按比例分层。此外,在选择某种抽样方法和操作策略时,我们不仅需要通盘考虑它的总误差、费用和可行性问题,而且要把它与其他可能选择的方法联系起来加以考虑。不仅如此,可供我们选择的抽样框常常会对抽样方法的选择有所限制。

在决定选择何种抽样方法之前,我们首先应该考虑等概和不等概选择各都有什么优点。选择的等概性保证总体的所有成员都有相等的被选入样本的可能性。样本是自加权的。在多级样本中,总的选择概率决定了选择的概率是否相等。

在不等概选择中,研究总体成员的选择概率是已知的,但是不相等的。不等概法需要用加权对那些因为对某些群体做了过度抽样而引起的偏倚进行修正。正因为如此,它会使估计值和抽样变异性的

计算变得更加困难。对于那些数据要满足多种用户的需要的调查而言，不等概选择所要求的加权和特定的计算方法势必会给他们带来诸多的不便。然而，不等概选择往往是提高估计的精度，或是使我们能对子总体做可靠的估计所必需的。

例如上面提到的让精神病人出院的研究采用的便是等概选择。从该调查用作抽样框的清单看，并不存在什么明确的重要的子群体。此外，调查数据将会有若干个人用不同方法进行分析。正因为如此，我们必须使用等概的选择方法。相反，在高龄老人研究中，我们则必须使用不等概的选择方法，以保证 11 个服务区的每一个至少都能有 100 个被调查者。此外，该研究还需要得到一些比较可靠的分地区的估计值，如果使用等概选择法，那些人口较少的地区在样本中的观察数就可能太少。我们要用权对那些因有比例地从人口较少的地区过度抽样所产生的偏倚进行纠正。在这样的情况下，加权的所失肯定低于我们地区估计值之所得。

在权衡选择各种抽样方法时，我们应该牢记对选择的概率问题做出的决策。选择和实施每一种方法所需要的资料，每种方法的长处和不足，以及抽样变异性的计算等问题我们将在以下四节中介绍。

简单随机抽样

最简单明了的抽样方法是简单随机抽样。总体的每一成员都有相等的选择概率。事实上，每一个样本都有相等的被选择的可能性。通常，简单随机抽样用于简单易行是我们最为关心的问题的场合。在这样的场合，那些比较复杂的抽样法才具有的长处，如估计值更精确等，都不是我们主要考虑的问题。

为了选取一个简单随机样本，研究者需要一份包含研究总体的所有成员的完整清单。我们可以用这样几种方法来选择样本：①给总体的每一成员分配一个唯一的识别数码；②在随机数码表中选择一个随机的起始数；③随机数码表数字的位数应该等于识别数码中的最高位的位数；④选取总体中每一个其识别数码与选取的随机数对应的成员；⑤舍弃任何在总体中不存在与之对应的识别数码的随机数；⑥不断重复这一过程，直至期望的成员数(n)完全选出，图 6.1

形象地展示了这一过程。对那些左边的数字有较小的数的总体,如本例中1 467的1 那样的总体,就可能有许多被舍弃的随机数。尽管这一问题可以用一个数学公式来避免,但这种做法是不可取的,因为它是以牺牲这一过程的简单易行的特点为代价的。

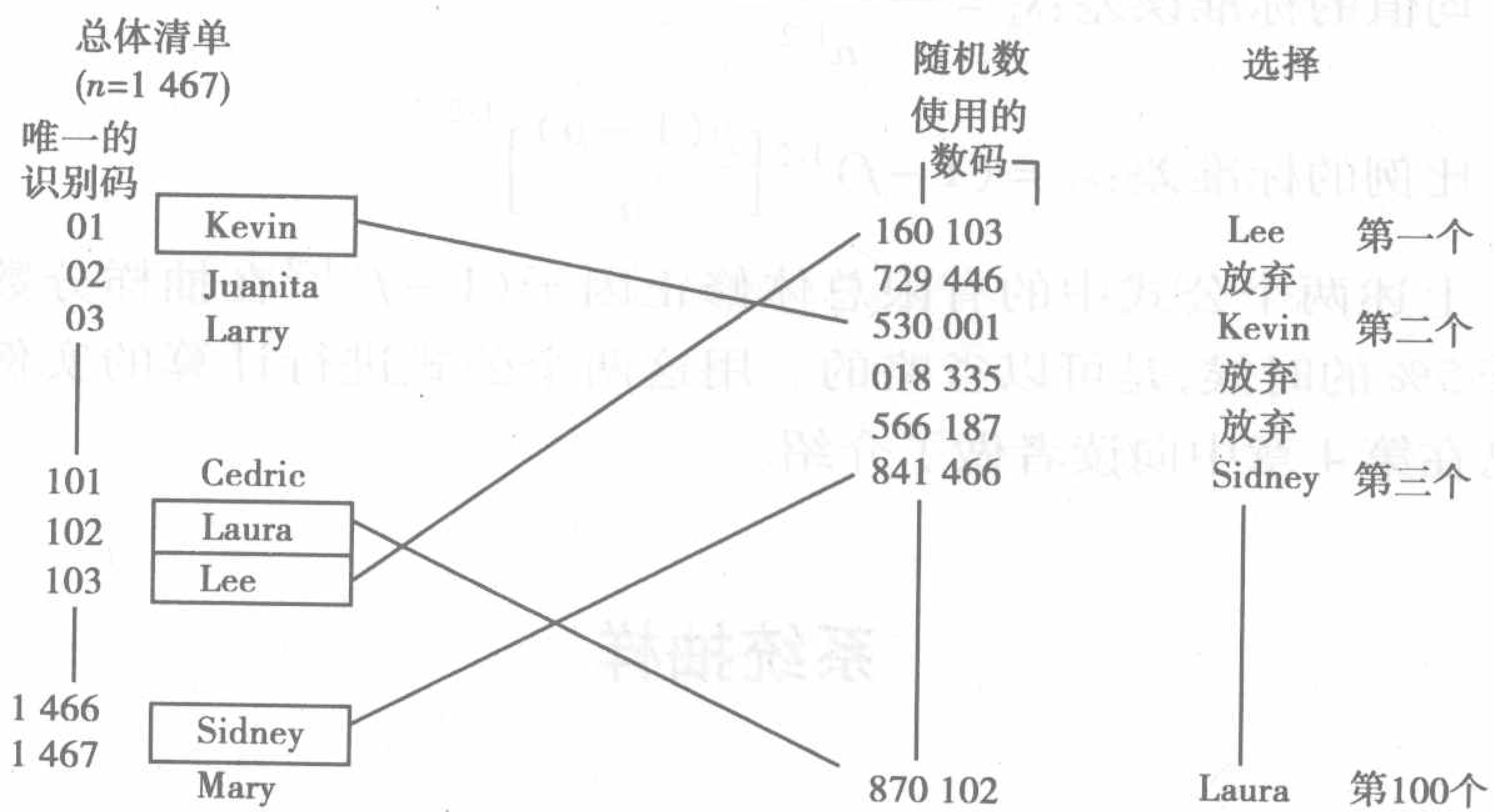


图 6.1　使用随机数码表,$n=100$

许多统计软件都有用以生成随机数,或从一份自动生成的清单中实际选取一个简单随机样本的程序。BASIC 语言是一种简单易用的计算机语言,大多数微型电子计算机都用它来编制生成一系列位于总体标志码范围之内的(伪)随机数(Baker, 1988, p. 147)。许多大型计算机和个人计算机的程序中,都有一个生成位于标志码范围内的随机数的子程序。

简单随机抽样的长处在于样本选取和数据使用的简便。大多数规范的统计软件在标准误差的计算时,都假定采用的抽样方法是简单随机抽样。因此这些程序在使用时都不必做任何修正或重新计算。在抽样框制作完成之后,不再需要有任何其他有关总体的信息便可进行抽样。

这种方法的不足之处也许正是其他方法具备的长处。这种方法需要一个明确的抽样框,也即一份列有整个研究总体的清单。如果数据收集过程涉及实地调查或个别访谈,那么简单随机抽样就会使调查散布在总体的地理范围所及的各个地点。如果研究的范围是全国性的或全州性的,那么交通费用必定相当高。最后,就每一抽到的单位的标准误差而言,这种方法的效率不如那些使用分层抽样的方

法高。

简单随机样本的连续变量和二分变量的均值的标准误差的计算公式如下式所示：

均值的标准误差：$s_{\bar{x}} = \frac{(1-f)^{1/2}s}{n^{1/2}}$

比例的标准差：$s_p = (1-f)^{1/2}\left[\frac{p(1-p)}{n}\right]^{1/2}$

上述两个公式中的有限总体修正因子$(1-f)^{1/2}$在抽样分数(f)小于5%的时候，是可以省略的。用这两个公式进行计算的实例，我们已在第4章中向读者做了介绍。

系统抽样

系统抽样具有与简单随机抽样类似的统计学性质。不过在某些场合，它能给研究者提供比简单随机抽样更多的便利。它给研究者带来的诸多便利之一是，便于在实地进行样本的选择。借助以下这些步骤：①确定样本的大小n；②确定选择间隔i，i等于N/n的商数去掉小数部分之后的整数；③设法得到研究总体的清单或具体代表物（如文件、票据）；④在1和i之间选择一个随机的起始数，以选取样本的第一个成员；⑤将每个位于起始数加上i的倍数处，即在$r+i$，$r+2i$，$r+3i$，…，$r+mi$处出现的总体成员选入样本，直到清单的末尾为止；⑥去掉那些在由步骤⑤中的随机选取过程选出的大于n的单元号码。

例如，假如我们确定需要从一个大小为15 222的总体中，通过有关票据的选择抽取一个容量为300的样本。将50.74（15 222/300）的小数部分舍去，得到抽样间隔50。从某一张表中选取18作为随机起始数。在完成上述步骤之后，我们从总体中的第18张票据开始选取样本，然后是总体中的第68和第118张票据等，如此这般，一直选下去，直至总体中的票据全部选完。如果最后选出的成员数多于所需的300个，我们可以用随机的方法去除那些多余的成员。用这样一种方式选取样本有几个优点。首先，如果有某种总体的具体的代表物存在（如文件或票据），我们不一定必须要有一份总体的完整的

清单。第二,选取的过程并不一定需要使用随机数码表,或在实地使用的随机数生成器。如果事先给定了抽样间隔数和起始数,那么经过适当的培训,只要有一定文化程度的工作人员便可以胜任诸如这样的抽样工作。

系统抽样还有另外一个优点,那就是它可以用某些实际情况对总体进行分层,从而确保总体的某些特征的比例的代表性。为了达到事实分层的目的,必须在抽样之前,对总体的单元按用以分层的变量之值排列。不言而喻,这时我们必须掌握整个研究总体的用以分层的变量的信息。用以分层的变量既可以是连续变量,譬如年龄,也可以是离散变量,如学生的成绩等级。这样的变量被我们用于对总体进行分组,以使样本在分层变量上能成比例,有更好的代表性。

精神病患者出院研究这一例子便用了事实分层的方法。该研究对从六个公共精神病院转到社区治疗机构的病人进行了调查。该研究先按病人出院的医院对登录的病人进行了分组,然后系统地选取样本,从而使医院的出院病人数的比例有代表性。

在进行系统抽样时,我们大家要注意一点。那就是,抽样框中的单元必须是完全混合的,或按分层的要求有目的地排列的。如果我们在不经意间使抽样框的排列具有某种周期性,就有可能使样本产生某种偏倚。例如,某一个样本框是按时间排列的月销售情况,选择的抽样间隔是 6,且设计规定只需要从 12 个月中选取 2 个月,假如我们最终选出的是三月和九月这两个月。如果情况果真如此,那么数据中存在的季节性模式便会被忽略,从而使结果产生偏倚。为了避免诸如这样的偏倚的发生,我们可以把抽样间隔缩小为 5。但这时有可能最终会选出三个月份,超出了设计规定的 2 个。如果这种情况果真发生了,我们只需要像前面介绍的那样,舍去多余的个案就可以了。

在抽样工作需要在实地进行时,这一方法会有一个不足,它会使实地工作的地点遍布于研究总体涉及的所有地理区域。调查地点过于分散会使交通费和支持实地工作所需的其他费用高到难以承受的地步。计算简单随机抽样的标准误差的公式也可以用于系统抽样。

分层抽样

分层抽样需将研究总体组合成一个个层，然后再在每一层选取一个随机样本。使用分层抽样的目的不外乎以下几个：确保每一层的比例有代表性；降低抽样变异性；使较小的子总体能在样本中有足够的数目，从而使分析更为可靠。

分层抽样可以是成比例的，也可以是不成比例的。成比例的分层抽样的目的在于在每一层中用相同的抽样分数来保证层的比例的代表性。表6.1显示了一个比例分层抽样的实例。在这一例子中，我们用母亲的文化程度将学生分成了三层（组）。这一变量被认为是认知技巧的一个重要的预测变量，因而保证样本能在这一变量上按比例地代表三个组是很重要的。这三个组是高中以下，高中或大学肄业，大学。

表6.1　按母亲文化程度分层的考试分数比例分层样本

	学生成绩				
A1层		A2层		A3层	
高中以下		高中和大学肄业		大学	
1	96	1	120	1	121
2	103	2	107	2	132
3	113	3	99	3	154
4	91	4	127		
5	122	5	139		
6	107				
7	103				
n_h	7		5		3
$\bar{x}_h$	105.0		118.4		135.7
s_h	10.34		15.87		16.80
N_h/N	0.467		0.333		0.200

n_h 是层样本量，
$\bar{x}_h$ 是层均值，
s_h 是标准差，
N_h/N 是以层的比例为根据的权数。

比例分层抽样。在使用比例分层抽样时，因为每一层包含的成员数目不同，所以每一层的样本的容量，如表 6.1 所示也不同。对于比例分层抽样来讲，每一层的抽样分数都是相同的。它的均值、比例和其他统计量的计算公式与简单随机抽样都是一样的。例如：

$$\bar{x} = \frac{\sum x_i}{n}$$

不过它的计算标准误差使用的公式却与简单随机抽样有所不同：

$$s_{\bar{x}} = \left(\sum \frac{w_k{}^2 s_k{}^2}{n} \right)^{1/2}$$

为了计算标准误差，这一公式增加了层的权数（$w_k = N_k/N$）。在比例分层抽样中，N 可以取代公式中的 n。该公式用总体中层的相对大小来为每一层的标准差的平方（$s_k{}^2/n$）加权。表 6.2 便是一个用这一公式进行计算的实例。

表 6.2　分层样本标准差计算法：以比例分层样本为例

	母亲文化程度层		
	A1	A2	A3
w_h	0.467	0.333	0.200
$s_h{}^2/n_h$	15.29	50.36	94.11

$s_{\bar{x}}{}^2 = (0.467)^2(15.29) + (0.333)^2(50.36) + (0.200)^2(94.11)$

$s_{\bar{x}}{}^2 = 3.33 + 5.59 + 3.76$

$s_{\bar{x}}{}^2 = 12.68$

$s_{\bar{x}} = (12.68)^{1/2} = 3.56$

式中，w_h 是每层的权数（N_h/N），

$s_h{}^2$ 是层的方差，

n_h 是层的样本量，

$s_{\bar{x}}$是均值的总标准误差。

续表

设计效应(Deff)
标准误差计算法:假设是简单随机样本 $\bar{x}=115.6$ $s=17.42$ $s_{\bar{x}}{}^2=17.42^2/15=303.40/15=20.23$ $s_{\bar{x}}=(20.23)^{1/2}=4.50$ 式中,$\bar{x}$是总均值, s是标准差, $s_{\bar{x}}$是均值的简单随机样本的标准误差。 $设计效应=\frac{12.68}{20.23}=0.63$ $(\mathrm{deff})^{1/2}=\frac{3.56}{4.50}=0.79$

分层可以减少标准误差,这一点可以用这一例子的未作分层数据,且使用用于简单随机样本公式计算的标准误差得到证明。分层样本的标准误差与简单随机样本的标准误差的比率就是设计效应(deff)的平方根,设计效应的值列在了该表的底部。就这一例子而言,精度的相对增益是21%(100% -79%)。精度增益的量取决于两个因素:

- 层间差异性(Variability Between Strata):各层均值与整个均值的差别越大,精度增益就越大;
- 层内同质性(Homogeneity Within Stratum):层内相似程度越高,精度增益就越大。

凯思(Kish ,1965)用为分层变量解释的研究变量的方差(R^2)的概念来解释这一问题。分层变量的解释力(explanatory power)越大,依据这一变量所做的分层的精度增益就越大。增益与被解释的方差的量成比例。表6.3将备择的分层变量班级用于与表6.2相同的总体。这一分层的增益为2%,它等于设计效应的平方根,小于用母亲的文化程度分层的增益。班级分层的设计效应的平方根为98%,而用母亲的文化程度分层的是79%。母亲的文化程度较之学生就读的班级解释了更多的学生考试成绩的方差。比较这两个例子,我们不

难发现,用母亲文化程度分层的各层均值的间距(30.7),大于用班级分层的均值间距(16.0)。不仅如此,我们还可以看到,用母亲文化程度分层的层内变差或标准差(s)都比较小。

表 6.3　分层样本标准误差计算法:以备择比例分层为例

	班级层		
	B1	B2	B3
	96	103	113
	107	120	99
	139	121	132
	127	103	154
	122	91	120
n_h	5	5	5
$\bar{x}_h$	118.2	107.6	123.6
s_h	16.90	12.76	20.77

$s_{\bar{x}}^2 = (0.333)^2(57.14) + (0.333)^2(32.56) + (0.333)^2(86.26)$

$s_{\bar{x}}^2 = 19.55$

$s_{\bar{x}} = 4.42$

式中,$\bar{x}_h$ 是层均值,

s_h 是层标准差,

$s_{\bar{x}}$ 是均值的总标准误差。

设计效应(Deff)

$$设计效应 = \frac{19.55}{20.23} = 0.97$$

$$(\text{deff})^{1/2} = 0.98$$

凯思还进一步指出,层所占比例的大小(w)会对相对增益有所影响。增加某一小子总体的样本量,不论它与总体的其余部分有多么大的不同,也不会显著改善估计的精度。最后他还指出,在百分比或比例是分析的主要目标时,分层的增益一般都不会太大。在层内的变差较小(同质)和层间的差别比较大的时候,分层就会有比较大的斩获。一般讲,在层与层之间的大小比例差别很大时,我们很难通过分层来提高估计值的精度。

诚如以上所述,比例分层抽样的优点在于能提高估计值的精度

和确保分层的群体的比例的代表性。分层本身并不需要什么额外的费用。但研究总体的每一成员都必须列出，并按用于分层的变量分类。而要得到有关整个总体的诸如这样的信息的费用则可能十分昂贵。有时得到与我们期望的分层的变量有关的信息的费用则可能不那么昂贵。例如，从成本效益的角度看，收集有关整个学生总体母亲文化程度的信息可能得不偿失。但是我们却有学生居住的地区的信息，可作为社会经济地位的信息的指标加以利用，而这一指标可能与母亲的文化程度相关。

不成比例的分层。研究者在遇到整个样本的精度，或某个子总体的精度不够这样的情况的时候，可以改而采用不成比例的分层。不成比例的分层源于对不同的层使用不同的抽样分数。采用不同的抽样分数将导致选择的不等概及最终样本中的代表性不成比例。为了修正选择的偏倚，加权是必不可少的。

不成比例的分层好处在于增加某一有较高的标准差的层的抽样数，而使该层的抽样变异有所降低。理解这一做法为什么能降低变差这一点，将对我们理解和掌握标准差的计算公式不无帮助。

$$s_{\bar{x}} = \left(\sum w_k^{\,2} s_{\bar{x}k}^{\,2} \right)^{1/2} = \left(w_1^{\,2} s_{\bar{x}1}^{\,2} + w_2^{\,2} s_{\bar{x}2}^{\,2} + \cdots + w_h^{\,2} s_{\bar{x}h}^{\,2} \right)^{1/2}$$

式中，$s_{\bar{x}k}$是第 k 层的标准差，而 w_k 则等于 N_k/N。

除了每一层的标准差已经计算出来之外，这一公式都与上面介绍的比例分层的公式相同。这两个公式的任何一个，既可用于比例分层，也可用于不成比例的分层。

因为我们必须先计算每层的标准差，然后再把它们合并成一个加权平均数，所以标准差最大和权数最大的层对标准差的影响最大。用不成比例的分层增加具有最大的抽样误差的层的层内样本量，将会降低该层的抽样误差，从而使整个样本的抽样误差也有所降低。表 6.4 用不成比例分层对这一性质做了阐述。该表的分层与表 6.2 和表 6.3 中的例子相同。用设计效应的平方根这一指标，我们测得抽样误差降低了将近 2%，每一种不成比例的分层的抽样误差都小于对应的每一种比例分层抽样误差。例如按母亲教育程度分层的高效单元分配(the efficient allocation of units)的标准误差为 3.48，而以同一变量作比例分层的则为 3.56。前者略低于后者，并不是特别显著。但是如果在这些例子中，当总样本量大于15 的时候，不成比例抽样

对总抽样误差降低的作用便会进一步显现出来。

表 6.4 分层样本量高效分配：层样本量变动的结果

母亲文化程度层	A1	A2	A3
$\bar{x}_h$	105.0	118.4	135.7
s_h	10.34	15.87	16.80
n_h	5	6	4
$s_{\bar{x}}=3.48$			
$(\text{deff})^{1/2}=0.77$			

班级层	B1	B2	B3
$\bar{x}_h$	118.2	107.6	123.6
s_h	16.90	12.76	20.77
n_h	5	4	6
$s_{\bar{x}}=4.34$			
$(\text{deff})^{1/2}=0.96$			

式中，$\bar{x}_h$ 是层均值，

s_h 是层标准差，

n_h 是层的大小，

$s_{\bar{x}}$ 是均值的总标准误差。

为了最大限度地提高总体估计值的精确度，样本量应与标准差和层的大小成比例：

$$n_k = \frac{n(N_k S_k)}{\sum N_k S_k}$$

式中，n_k 是层的样本量，

n 是总样本量，

N_k 是总体的层的容量，

S_k 是层的标准差。

虽然这一公式看起来似乎是合乎逻辑的，但实际上总体和层的标准

差几乎都是未知的。所以，分配给每一层的抽样单元不是那么精确，但是在标准差，或在更多的时候是它们的相对大小可以被估计的时候，对层的抽样单位数进行分配将会使估计值的精确度有所提高。在表6.4所列的两个例子中，我们对第一个例子中的15个抽样单位中的2个做了重新分配，对第二个例子中的一个抽样单位作了重新分配。在第一个例子中，我们将A1层中的抽样单位减少2个，给A2和A3分别增加了1个。在第二个例子中，我们从B2层中取走了一个抽样单位，加到了B3层。我们可以用与估计高效样本量的标准差相同的方法来估计每一层的变差。这些方法有事先研究、探索性研究、使用值域(using the range)等。

还有一种场合也需要采用不成比例的抽样，那就是在遇有我们需要对子群体进行分析，而按比例选取产生的子样本的标准差又过于大的时候。这时不成比例的分层抽样使我们得以在不必成比例地加大总样本量的前提下，加大子总体的样本量。要能这样做，我们要以能使子总体的成员与这一特定层次联系在一起的方式来定义这个层。实现这一目的的理想分层都发生在层是由互斥的子总体的成员组成的时候。在子总体的成员是高度集中的时，不成比例抽样也同样可以使用，但效率可能会有一定损失。

不成比例分层的主要缺点是在计算标准差的时候必须要加权。这样标准差的计算势必会更加复杂。此外，保存的数据集中不仅必须有专门用来识别层的编码，还必须包括分层赖以生成的权数。许多统计软件都有用于设置总体估计值和标准差计算的权数的程序或命令。抽样中最为常见的错误是在样本选择时采用的是不成比例的抽样，但却在估计过程中没有加权，因而未能对总体估计值中的这一偏倚进行修正。导致这一问题的主要原因不外乎这样两种：一是在制订研究计划时，没有把抽样和分析作为一个整体考虑；二是在做数据分析的时候，分析人员对抽样设计不甚了解。加权、设计结构的使用等问题将在本书的最后一章再一次进行讨论。

在第4章介绍的那些各具特色的抽样实例中，有三个例子使用了分层抽样。全美住户调查的第一级选择使用了74个层。每一层选取一个初级抽样单位。此外还有10个层，均是由肯定要选取的全美10个最大的大都市综合地区(SMSAs)组成的层。从每层选取一个抽样单位的设计给标准误差的估计造成了一定的困难。因为每层

只有一个单位，我们就难以估计层内的方差。这样的设计要求在标准误差计算的过程中，将那些最为相似的层组合在一起。

高龄老人调查用州内的 11 个地区作为层。研究设计规定每一地区抽取的单位不得低于 100，总样本量约为 1 500 个抽样单位。该调查采用了不成比例的样本分配，其原因有二：一是增加地区内和地区间的调查结果的可靠性；二是实行这样的样本分配方案将增加 4 个地区的样本量的配置比例，降低 7 个地区的样本量的配置比例。例如在 1 号地区，完成 113 个个案的调查，在加权估计时只相当于 33 个。因此，在加权之后，113 个个案便转而代表了样本的 2.0%（33/1 647 =0.02）（参见表 4.4）。

北卡罗莱纳州居民调查原来使用的是比例分层样本。用于编制抽样框的两种条目中的每一种都构成了一个层。每层的抽样单位数的分配以它在总体所占的比例为依据。然而，经验告诉我们来自医疗补助名单的回答人数，将会少于那些来自所得税的回答人数。为了既能使样本中来自医疗补助清单中的人数有所增加，又能得到一个可实际使用的自加权样本，我们必须对每一层的比例进行调整。在调整之后，医疗补助清单的比例由原来的 11% 提高到了 13.5%。

整群抽样

整群抽样是随机选择法的一种。这种方法选择的是被我们称作为群（clusters）的编组，选出的每一个群中的所有成员都将被选入样本。在可资利用的资料是一份群的清单，而非一份总体的清单时，整群样本是非常有用的。这样的情况常常发生在一个项目在若干地方或地区性管理机构管辖范围之外的地方实施，而在调查时又缺乏近期的客户汇总资料时候。在数据收集涉及实地访问，或从地区或地方性的办公机构得到一些有关记录的时候，整群抽样也非常有用。在这些场合，整群抽样会大大降低交通费和培训费。

尽管整群抽样有比较经济实用的优点，但与此同时，标准误差会有所上升，因为它会使选择的独立性有所下降。有鉴于此，在考虑选择使用整群抽样的时候，我们需对这二者做一番权衡。在简单随机抽样中，每一个抽样单位的选择都独立于其他抽样单位的选择。但整群样本却不然，每一个群的选择都是随机的，因而也是独立的，但

是每一个样本单位的选择却不是独立的。这就是说,样本单位之所以被选入样本,取决于群的选择。这样一种选择方式将会导致选择的独立性的丧失。而每一抽样单位的独立性信息的丢失将会导致精度的丢失。

信息丢失造成的影响可以在用于估计整群样本的抽样误差的公式中得到证明。在这里,各个群是大致相当的。

$$s_{\bar{x}} = \left[\frac{(1 - \frac{a}{A}) \sum (\bar{x}_a - \bar{x})^2}{a(a-1)} \right]^{1/2}$$

式中,$s_{\bar{x}}$是标准误差,

a 是选取的群数,

A 是总体中总群数,

$\bar{x}_a$ 是群均值,

$\bar{x}$ 则是总均值。

公式中,选取的群数替代了样本的容量,而总体中的群的数目替代了总体的大小。这是因为选取的群的数目是独立选择的数目。此外,我们还必须指出偏差的计算式$(\bar{x}_a - \bar{x})^2$,是群的均值减去总均值。

公式中的第一项是有限总体修正式。在选取的群数即将穷尽可供选取的群的数目时,这一修正式将会对整群样本的标准误差进行修正。表 6.5 是一个计算的实例。群的均值是 105.0,118.4 和 123.4。群的均值与总均值(115.6)的偏差的平方是 30.17。这一数目的平方根是5.49。表中同时列出了假定用简单随机抽样得到的同一样本的偏差的计算结果。整群样本的设计效应约为 1.49。这就是说,整群样本的标准差是同容量简单随机样本的 1.22 倍,这一数字就是设计效应 1.49 的平方根。

设计效应的大小取决于三个因素:

- 群的均值与总均值之间的差别。
- 群的异质性。
- 选择的群的数目。

前两个因子是密切相关的。标准误差随着群的均值与总均值之差的加大而加大。因为在式中,我们取了这个差的平方,从而使公式对差的敏感性有所增加。因此,群的差异越大,估计值的精度就越低。

表6.5 整群样本标准误差计算法

	群 1	群 2	群 3
	96	120	103
	103	107	121
	113	99	107
	91	127	132
	122	139	154
$\bar{x}_a$	105.0	118.4	123.4

$\bar{x} = 115.6$

$s_{\bar{x}}^2 = \sum(\bar{x}_a - \bar{x})^2/(a-1)a$

$s_{\bar{x}}^2 = [(105.0-115.6)^2 + (118.4-115.6)^2 + (126.0-115.6)^2]/(3\times 2)$

$s_{\bar{x}}^2 = 30.17$①

$s_{\bar{x}} = 5.49$

式中，$\bar{x}_a$ 是群均值，

a 是群的数目，

$\bar{x}$ 是总均值，

$s_{\bar{x}}$是均值的总标准误差。

设计效应(Deff)

标准误差计算法，假设是简单随机样本

$\bar{x} = 115.6$

$s_{\bar{x}}^2 = 20.23$

$s_{\bar{x}} = 4.50$

设计效应 = 1.49

$(\mathrm{deff})^{1/2} = 1.22$

在群是同质的时候，均值差就会比较大。在每群的标准差都与总体的标准差相等的时候，设计效应等于1，说明整群样本的标准差与简单随机样本相同。群内差异越大，样本精度越高。最后要说明的一点是，群的数目的多少也会对样本估计值的精度有影响。这一点与简单随机样本中增加样本量对样本估计值的精度影响颇为相似。标准误差

①这里的计算公式中遗漏了有限总体修正式 $1-\frac{a}{A}$，所以按公式计算的 $s_{\bar{x}}^2$ 并不等于30.17。——译者注

的变化如同一个群的数目的平方根函数。增加群的数目可以提高样本的精度。在采用整群抽样方法的时候,选取更多的群间变差较小的群便会使精确度有所提高。不过增加群的数量必然会导致收集数据费用的增加。不言而喻,已经为我们所熟悉的,如何在费用和精度之间作出合理的抉择这一问题,又一次摆在了我们的面前。

分层可以在一定程度上消除使用整群抽样导致的标准误差加大的影响。从理论上讲,这是因为将层内群的加权均值的标准差合并可以提高样本的精度的缘故。此外,通过选择代表各种特质的群,我们还可以提高样本的信度。

一般我们常将地理群体或原群体(intact group)作为群来使用。学校,或作为学校的备择对象的班级,常在与教育问题有关的调查中作为群来使用。在管理或评估研究中,地区性的机构或诊所也有可能作为群来使用。一旦我们使用了诸如这样的群,这就意味着研究者已经无法再提供已经选取的单元之外的,其他特定地区或学校的信息了。在有关政治问题的调查中,如果排除某些地区,就有可能损害样本的可信度。

多级抽样

比整群抽样使用更多的抽样方法是多级抽样,它与整群抽样颇为相似。比较简单的多级抽样设计是二级抽样。它将群的选择作为初级选择,然后再在选出的群中抽取成员,产生最终的样本。更为复杂的多级抽样需要进行多次抽样单位的选择,涉及大群套小群的选择,即先抽取大群,然后再抽取大群内的小群,更小群,直至在最小的群中抽出基本的单元。全美住户调查采用的就是多级抽样方法。

密歇根大学社会调查研究所的调查研究中心(SRC)实施的全美住户调查采用了五级抽样。如果将收集个人层次上的数据也作为一级,那么就有六级。这一调查是一般人口总体面积概率抽样(area probability sampling)的一个范本。各级中的每一级都以面积定义。第一级是大都市或县的区域。随后各级的区域逐级变小。在第二级,我们将从一张由已经选出的初级抽样单位组成的全部单位的清单中,抽取一定数目的市、镇或农村地区。然后再继续这一过程,三

级、四级……直至从一个含有四个住户的群中抽出一或几个住户。

在这一样本设计中，我们在组成美国本土部分的 2 700 个初级抽样单位中，抽取了 74 个抽样单位，然后从每一个初级单位中平均抽取五个二级抽样单位。这样二级抽样单位的总数为 370 个。再通过第三和第四级抽样，最终达到片（一种由四个住户组成的单位）的抽取。抽取片的过程，城市和乡村有所不同。在城市里我们先抽取街区，而在农村地区则先抽取“地块”，然后再从抽出的街区或地块中抽取片。地块和片的选取需要借助每一城市、镇和农村地区的地图。二级抽样单位被分割为含有 16 个住户的群，然后再把群分割成在第四级要抽取的片。住户是最后一级的抽样单位。

构成多级抽样的基本原理是简单明了的。然而设计的实施却需要有相当高的专业水准，否则无法避免在不经意间产生的偏倚。例如，如果预定某一个群应该含 16 个单位，但实际上该群的所在地正在发展成一个有 200 个住宅单位的公寓群。如果我们不增加这一个群的样本量，或通过加权对选择的概率进行补救，那么样本就可能会有偏倚。不言而喻，为了避免样本的偏倚而要作的决策和对实际的操作过程进行监督以得到决策需要的信息，都需要在面积概率抽样方面具备相当的实际经验和专业知识。

多级抽样的一种更为易于操作的用途是通过嵌套单位的选择，得到一个特殊总体的样本。例如，研究者可能需要抽取一个高中生的样本，以确定他们对其他种族和文化的人们的态度。在第一级，我们可以抽取某一大都市地区内以地区、大小和地理位置分层的校区。在第二级，则在选中的校区内抽取班级，这些班级既有职业高中的，也有普通高中的。最后，我们在选出的班级中，抽取学生作为我们的被调查人，如图 6.2。

在使用这一基本设计时，研究者既可以采用等概选择法，也可以采用不等概选择法。而究竟应该采用什么样的选择法，则要视具体的研究目的而定。假如我们想要对总体中的一小部分成员，例如少数民族学生，进行子总体分析，那么我们可能就应该使用不等概的不成比例的样本。当一种可供我们选择的，使用比较普遍的等概样本选择法，它的初级单位选择概率与其单位的大小或比例（PPS）。

为了根据 PPS 抽取校区，一份如表 6.6 那样的含有一个单位大小的量度的（在本例中，它就是高中生的数目）的清单是不可缺少的。

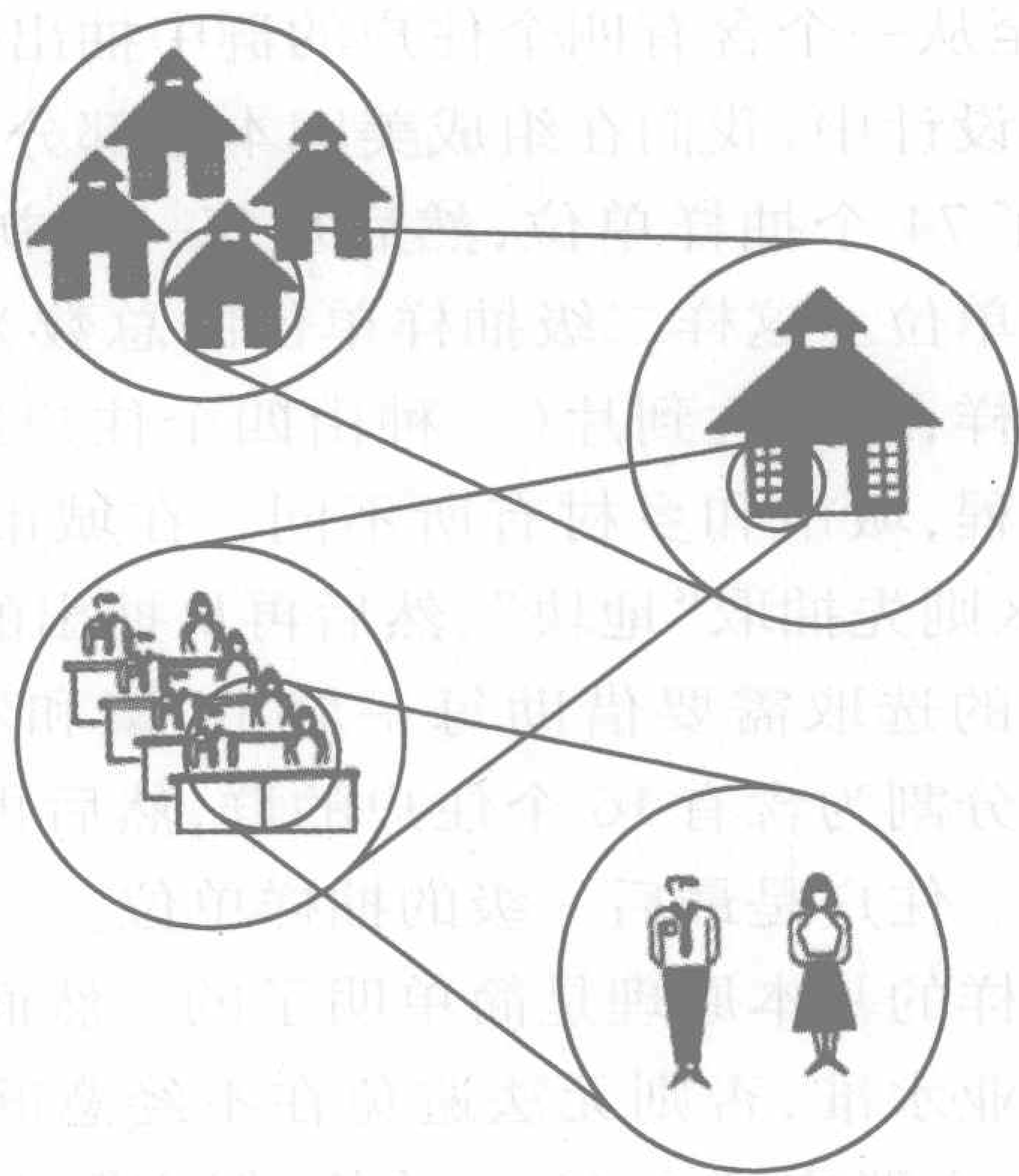


图6.2 多级抽样

在得到了这样的清单之后，我们必须先选取一定数目的初级单位。在这一例子中，因为交通费用有限，所以我们只选取了30个校区。在校区选出之后，我们还需要计算抽样间隔。抽样间隔等于学生总数除以单位数(325 000/30 = 10 833)。在抽样间隔确定之后，我们还需在1到10 833之间确定一个随机的起始数(r)。在u和r都确定之后，初级单位便也就确定了——它就是那些含有r和$r+10\ 833u$个学生的校区，式中的u等于1到29(原文意思不明确，译文根据原意做了修改，译者)。

在用表6.6的有关资料进行抽样时，如果我们选择的随机起始数是7 278，那么第一个选中的单位便是列在表中第二位的Roanoke。第二个选出的单位就是那个含7 278 + 10 833 = 18 111那个数的单位。这就意味着Amherst和Nottoway这两个单位将被跳过，而Fairfax County则成为了第二个被选出的单位。然后按这样的方法，再选第三、第四个，直至最后一个单位。而最后一个选出的单位应该是A-melia，因为它含有累计的第321 435[(10 833 ×29) +7 278]个学生。

任何一个校区的选择概率都是$N_d/10\ 833$，而N_d则代表校区的大小。将初级单位选择法和5个班级及每个班级10个学生的方法合在一起，我们将为样本中的每一个学生提供0.004 5的总选择概率。概率究竟能在多大程度上相等取决于大小量度的精确性。班级

数目和每班学生的数目的乘积必须等于高中学生的数目(大小的量度)。总样本数约为 1 500(30×5×10)。

在使用多级抽样时,研究者可能希望预先确定一些进入样本的初级单位。如研究者可能希望将一些主要的城市,如伊利诺斯州的芝加哥选入样本。实际上在表 6.6 这一例子中,那些有 10 833 个以上高中生的校区,如 Roanoke 和 Fairfax 肯定是都会被选入样本的。而在全美住户调查,预先确定进入样本的初级单位也有 12 个之多。

这种做法似乎在某种程度上违反了抽样的原则。当然,那些确定要选择(预先选择)的单位肯定是不能被看做抽中的单位的。但是,只要其余的各级进行的选择所产生的概率等于其他以等概设计选取的单位的概率,或能用加权对预先选择进行适当的补救,预先选择并不会导致样本的偏倚。让我们看一下作为例子的表 6.6 中各级单位的选取概率。Roanoke 因为有 11 538 个高中生而使它的第一级的选择概率等于 1.07 (11 538/10 833),因为概率大于 1,所以它的选取具有了确定性——它肯定会被选入样本。但它的最后两级的选择概率则比较低(0.004 3),从而对它在第一级选择具有的确定性做了补救。苏德曼对遇有在无法得到目标总体的大小估计值或估计值不够精确这样的情况时,使用 PPS 进行多级抽样的方法作了介绍,并提供有关的实例(Sudman, 1976, pp. 134-146)。

表 6.6　与大小成比例的(PPS)样本的校区清单

校区	学生数	累计学生数
Albemarle	2 318	2 318
Roanoke	11 538	13 856
Amherst	1 217	15 073
Nottoway	584	15 657
Fairfax	46 154	61 811
⋮	⋮	⋮
Amelia	3 121	324 071
Winchester	929	325 000
	325 000	

估计多级样本的抽样变异是一个复杂的过程。许多可供我们使

用的商业软件都有用泰勒序列线性化(Taylor series linearizations)估计抽样变异的子程序。例如统计分析系统软件(SAS)便有为从复杂样本设计中的得到的各种统计值估计抽样误差的子程序(Holt, 1977; Shah, 1981)。这些方法需要在初级选择的每一层和每一级的每一个抽样单位中做一个以上的选择。这种做法可能会使第一级的可能的层数减少一半,而使分层的精度增益有所降低。另一种估计抽样变异的方法是反复抽取样本,然后再把单个的样本结果合并在一起,得到抽样变异的估计值。在使用这一方法时,抽样变异可用如下的公式计算:

$$s_{\bar{x}} = \left[\frac{\sum (\bar{x}_i - \bar{x})^2}{k(k-1)} \right]^{1/2}$$

式中,$s_{\bar{x}}$ 是抽样误差的估计值,

$\bar{x}_i$ 是第 i 个子样本的均值,

$\bar{x}$ 是总均值,

k 则是子样本的个数。

然而,重复样本、重复实验的实际使用会使初级抽样单位的数目受到限制,进而使可以利用的层的数目也受到限制。这样,就会导致某些精度因为设计的低效而丢失。不仅如此,在过多地使用重复实验时,这一问题还会更加严重。但是如果重复实验的次数太少,抽样变异估计值的又会不太可靠。

为了克服上面提到的那些问题,我们改进设计了一种或可称为均衡重复试验的方法(the method of balanced repeated replications)。这一方法要求从实验使用的每一个层中选取两个初级抽样单位(PSUs)。例如,一个使用 12 个层的设计需要有 24 个初级抽样单位。每个在某一层内的 PSU 将被标上一个"+"号或"-"号。若用表 6.7 的第一行中标出的那些初级抽样单位,将会有一半样本被选中。我们再用在第一个半数样本中被略去的那些 PSU 选取补充样本。我们用来计算抽样变异的估计值的公式是:

$$s_b = \left[\frac{\sum (b_r - b_c)^2}{4K} \right]^{1/2}$$

式中,s_b 是要估计的统计值(均值、回归系数等),

b_r 是重复半数样本的统计值,

b_c 是补充样本的统计值,

K 是重复的次数。

这一计算公式既简单明了,又能适用于各种各样的统计值。与泰勒近似法相比,人们更乐于使用这种方法。不过,我们必须说明的是,它的确需要分别计算两倍于层数的估计值。诸如表 6.7 那样用以指导我们对各种大小的备择的层选择半数样本的表,读者可在苏德曼(Sudman,1976)和弗兰克尔(Frankel,1971)的著作中找到。

使用均衡重复实验法要求从每一层选取两个 PSU 这一条件可以放松。在从一个层只选取一个 PSU 时,这些层可以合并成含有两个 PSU 的层。然而,我们必须指出的是,层的合并是以抽样变异的高估为代价的。不过高估的量往往是微不足道的。

表 6.7　备择层样本量半样本选择法例示

子样本	选取的层对											
	1	2	3	4	5	6	7	8	9	10	11	12
1	+	−	−	−	+	−	−	+	+	−	+	−
2	+	+	−	−	−	+	−	−	+	+	−	+
3	+	+	+	−	−	−	+	−	−	+	+	−
4	+	+	+	+	−	−	−	+	−	−	+	+
5	−	+	+	+	+	−	−	−	+	−	−	+
6	+	−	+	+	+	−	−	−	−	+	−	−
7	−	+	−	+	+	+	+	−	−	−	+	−
8	+	−	+	−	+	+	+	+	−	−	−	+
9	+	+	−	+	−	+	+	+	+	−	−	−
10	−	+	+	−	+	−	+	+	+	+	−	−
11	−	−	+	+	−	+	−	+	+	+	+	−
12	+	−	−	+	+	−	+	−	+	+	+	+
13	−	+	−	−	+	+	−	+	−	+	+	+
14	−	−	+	−	−	+	+	−	+	−	+	+
15	−	−	−	+	−	−	+	+	−	+	−	+
16	−	−	−	−	−	−	−	−	−	−	−	−

注:这张表格直观地阐述了从 12 个层中每一个层的成对的成员中选取 1 个成员形成一个半数样本的过程。在 1 ~ 12 这 12 个层中,每一层都只有两个成员,一个被标以"+"号,而另一个则被标以"-"号。我们要从这 12 个有两个(一对)成员的层中的每一层选取 1 个成员形成 16 子样本,而余下的成员则构成 16 个与之一一对应的补充样本。例如表中的子样本 5,我们从第一个层对中选取的是标以"+"号的成员,而从第二层对中选取了标以"-"号的成员。我们先以这样的方式选取 16 个半数样本,计算每一个半数样本的将被用于标准误差估计的统计值(x,b 等),然后将求得的统计值与从它的补充半数样本得到的统计值做比较。最后,将计算 16 个子样本与其补充样本的差得到的结果合并起来,便得到了标准误差的估计值。

资料来源:Frankel,1971。本书引用已得到作者允许。

最后一种估计复杂样本的抽样变异的方法是折刀法(jackknife)。虽然这种方法也与重复实验有涉,但是在这种方法中,每次实验都要计算单独一层对抽样方差所做的贡献。折刀重复实验法要求从每一层选取至少两个PSU,但与均衡实验法不同的是,它不限于每层只取两个PSU。每次实验,都要移去一个PSU,同时对层中留下的其余的PSU加权以作补偿,并估计统计值。现在,我们能假定在目前这一例子中,每层抽取两个PSU,我们移去层中的另一个PSU,来计算补偿重复实验,进行加权,并估计统计值。单独一层的方差估计值将被合并在一起,以对总的抽样变异做出估计。

诚如这些计算方法所示,在我们做出采用多级抽样方法这一决定之后,我们还需要做出一些重要的决定。其中最为关键的是决定层和初级抽样单位的数目。层的数目越多,分层的变量与我们感兴趣的变量之间的相关程度越高,样本的精度就越高。层越多,需要做的重复实验的次数也就越多。因此,更多的层势必会增加抽样的复杂性,进而增加计算机计算所需要的费用。在需要使用最大似然或迭代求解的计算机程序时,计算费用的增加会特别显著(不过随着计算机性能的提高和统计程序的发展,这种成本已经大大降低。译者注)。然而必须指出的是,过少的层将会使样本变量的估计值变得不太可靠。赫斯认为,使用回归分析的多级样本能有30到50个层就相当不错了(Hess,1985)。

小　结

抽样方法的抉择取决于各种因素。表6.8列出了一些在作抽样方法问题抉择时,应该考虑的最为重要的因素。在深思熟虑之后,对某种方法作出了选择,这就意味着我们已经对如何进行实地操作(包括得到所需的记录或对被调查人进行调查的许可和设计制定抽样框等)、非抽样误差和抽样偏倚,以及所需费用等问题也已经有了一个大致的计划。这时,如果要使这一设计终将产生的误差控制在可以容忍的范围,那么样本容量便是我们接下来所应该考虑的问题。

表 6.8 抽样方法选择法

	简单随机样本(SRS)	系统	分层	整群	多级
抽样框					
一般	随机数码拨号(RDD)	编制小地理区域的清单	用分层变量编制小地理区域的清单	地理单位	面积概率样本;瓦克斯伯格 RDD
特殊	编制清单	物质代表或编制清单	用分层变量编制的清单	编制群的清单	编制初级抽样单位的清单
数据收集法	电话 RDD;其他任何适用的方法	任何适用的方法	任何适用的方法;使用不同的方法很有用	任何适用的方法;个别访谈和录音访谈很有用	面积概率个别访谈;电话 RDD;任何适用的方法
优点	简单易行自加权	不需要用做抽样框的清单;伪分层	提高效率;子总体分析	不需要基本单元的清单;集中进行个别访谈	与整群样本相同,但效率更高
费用	—	—	不成比例,需要加权,需要支付分层信息的费用	会增加抽样误差	会增加抽样误差,抽样误差计算花费颇多

第7章 样本容量

Sample Size

诚如前述，样本单位数，即样本容量往往是研究小组告诉抽样顾问的第一个问题。而问题的具体答案则要等待其他方面的设计和抽样方法的选择确定之后才能确定。为什么样本的容量对于一个调查研究竟会如此的重要呢？样本容量是我们使估计值达到决策和科学探索所要求的精确性和可靠性的有力工具。从图 7.1，我们不难看到增加样本容量对抽样变异的估计值的影响。向下倾斜的曲线表明抽样变异随样本容量的增加而减少。每增加一个单位的样本量，在样本容量较小时的精度增益，大于样本量较大的时候。

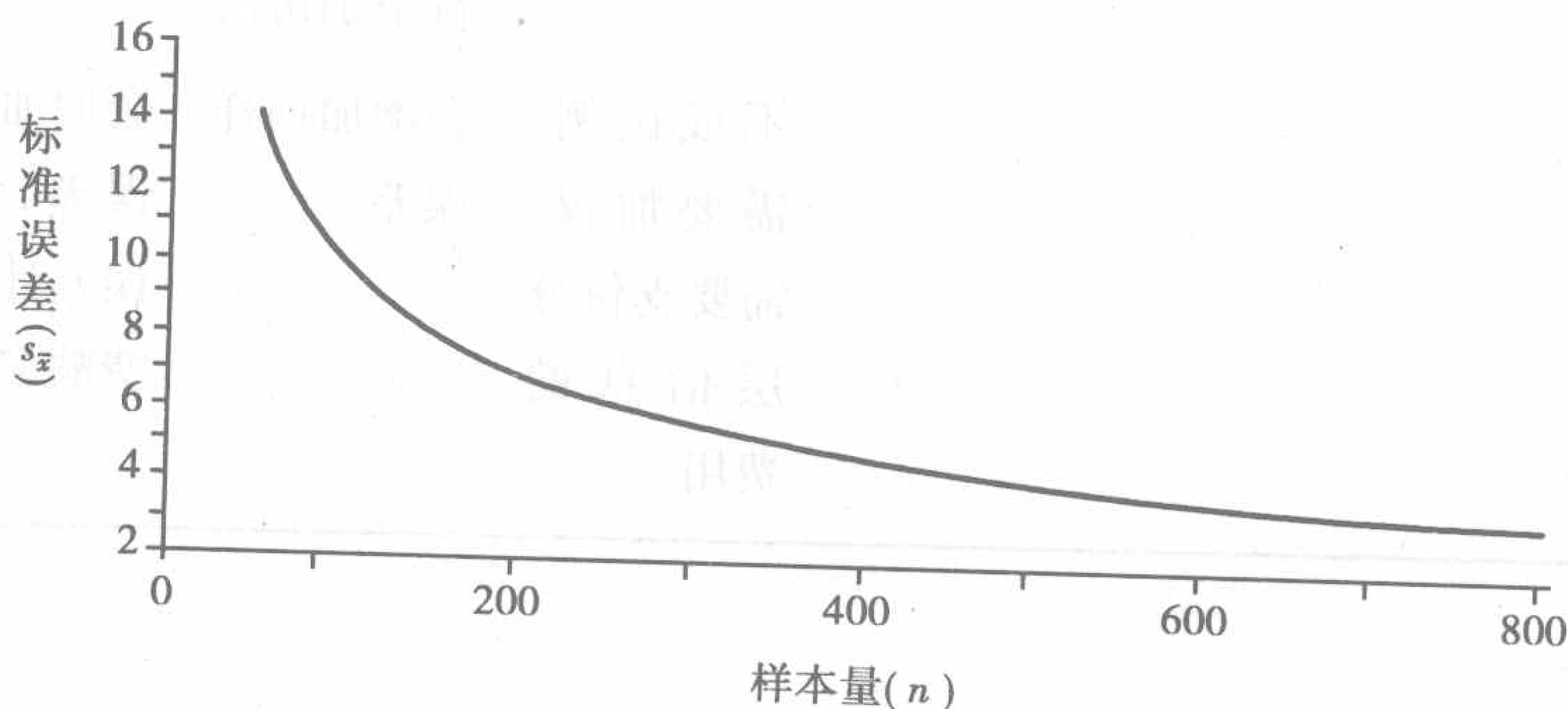


图 7.1　样本量和标准误差的关系

不言而喻，增加样本容量必然会使费用有所增加。较大的样本需要有更高的费用来进行数据收集，而在数据收集使用的方法是访谈的时候，情况尤其如此。此外对无回答者的追踪调查、编码和数据分析的费用，也必然会因为样本的加大而加大。例如在需要花费大量时间和精力去追踪那些无回答的被调查人时，增加的样本量就有

可能使总误差因非抽样偏倚上升而上升。我们不能凭空去考虑样本容量问题。在考虑样本容量问题时,我们必须再一次在费用、总误差和其他的设计抉择问题之间权衡斟酌。

在进入样本大小选择过程之前,我们必须有条不紊地对一系列因素进行一番考察。首先,我们必须确定估计值可以容忍的误差,或功率分析(power of the analysis)的大小。各种政策研究,尽管它们与理论检验取向的研究目的有所不同,但它们遵从的标准却并无二致。对政策研究而言,可容忍的误差或效率需求的确定,似乎更倾向于以手头掌握的特定环境的信息而不是规范的标准为依据。

在政策研究中,设法让政策制定者参与数据要求的精度的决策是很有用处的。我们可以用编写诸如"如果……将会怎么样"(what if)这样一种类型问题的方式,简单明了地把我们的决策转达给政策制定者。例如我们可以这样对他们说:"我告诉您,研究估计有 60% 的高龄老人需要我们给他们提供服务,但是我们认为比较有把握地说,需要我们提供服务的高龄老人可能在 50% 到 70% 之间。我们提供的信息差距是 20%,你觉得这样的信息有用吗?"而在另一项有关提高成绩差的学生的成绩的研究中,也许我们可以这样来向政策制定者说明这一问题:

> 你知道,那些成绩差的学生的考试分数位于标准化考试的第 38 个百分位处。考虑到这个项目的费用和我们可以采取的措施,你觉得通过这个项目把那些学生成绩提高到第 40 个百分位怎么样?或者再高一点,到第 43 个百分位怎么样?

只要将诸如这样的问题作一些变化,便可用于某一特定的政策研究。在考虑精度问题时,我们必须对所需的各种费用予以尽可能正确的估计,以便能清楚地了解费用和抽样变异之间的关系。

在选择样本容量时,我们必须考虑的问题包括:

- 高效样本的容量
- 高效样本设计的含义
- 样本容量和子总体分析设计的含义
- 不合格问卷和无回答的纠正
- 设计规定的样本量的费用

- 信度

下面我们逐项对这些问题进行讨论。

高效样本的容量

高效样本容量的确定,是以将抽样变异限制在我们要求的水平所需的样本量的估计值为依据的。对于一个主要目的在于描述的研究而言,可容忍的抽样变异的大小是以估计值所需的精度而定的。对于分析性研究而言,高效的容量是使我们能探测到研究估计值的容量。尽管研究涉及更多的信息,但这些信息常常会反复使用,但一般讲,高效样本量的估计值仍然是基于简单随机样本这一假定的,所以特定设计中的高效样本量的估计值还是可以求得的。

描述性研究的高效样本量的计算开始于对可容忍的误差(te)的思考:标准误差乘以选定的置信区间的 t 值。必须对变量的方差或标准差进行估计。通常我们可以从以往的研究中来估计标准差。如果目前的目标总体与以往的目标总体有所不同,那么我们可能必须对它做一些修正。

我们也可以通过小规模的探索性研究来估计标准差。有的时候只需50个左右的个案便可为我们提供很有用的估计值。在采用这一办法时,我们最好能随机地从目标总体中选择个案。不过并非任何时候,都能如我们所愿的那样随机地选择我们的调查个案。我们必须对期望的总体估计值是否与探索性的总体研究得到的值有什么不同这一问题做出判断。如果有所不同,那么就需要做一定的修正。

第三种求标准差大致的估计值的方法是将值域除以4。当然这样做的前提是数据或专家能告诉我们变量的最高和最低值。我们可把值域估计值代入公式,进而求出标准差的估计值。

最后一种常用的估计方差的方法,大多用于研究者感兴趣的问题是变量的各种比例的时候。这一方法只要简单地假定在 $p = 0.5$ 的时候,方差达到最大就可以了。这时 $p(1-p)$ 的积便等于0.25,即一个比例或一种最差的情况(worst-case scenanio)的最大可能积。

均值和比例的高效的样本容量的计算公式及其计算实例已经在表7.1中列出。

表 7.1 均值的高效样本量

$s = 37.6$ $N = 22\ 000$

$te = 1.764$ $t = 1.96$

$s_{\bar{x}} = te/t = 0.9$

$n' = (37.6)^2/(1.764/1.96)^2$

$n' = 1\ 745$

$n = 1\ 745/[1 + (1\ 745/22\ 000)]$

$n = 1\ 617$

式中，s 是标准差的估计值，

N 是总体的大小，

te 是可容忍的误差，

t 是希望的置信水平的 t 值，

$s_{\bar{x}}$ 是允许的标准误差，

n' 是未作有限总体修正的样本量，

n 是作了有限总体修正的样本量。

比例的高效样本量

$p = 0.6$ $1 - p = 0.4$

$te = 0.02$ $N = 1\ 100\ 000$

$s_p = 0.01$ $t = 1.96$

$n' = (0.6)(0.4)/(0.02/1.96)^2$

$n' = 2\ 305$

$n = 2\ 305/(1 + 2\ 305/1\ 100\ 000)$

$n = 2\ 300$

式中，p 是某一样本比例，

te 是可容忍的误差，

N 是总体的大小，

s_p 是该比例的标准误差，

n' 是未作有限总体修正的样本量，

n 是作了有限总体修正的样本量。

在第一个计算实例中，标准差估计值是 37.6，可容忍的误差是 1.764，它与 0.9 个单位的标准差和 1.96 个 t 值相对应。在未做有限总体修正（FPC）前，高效样本的容量是 1 745，修正之后为 1 617。在

第二个例子中,可容忍误差是2%,它会产生 1% 的标准误差。这就是说,研究者有 95% 的把握确信比例的估计值将围绕真的比例值上下 2% 波动。为了得到这样的精度,在做了有限总体修正之后,我们需要的样本容量是 2 300。

尽管这些计算是比较简单易行的,且许多初级统计学教科书中一般都有高效样本容量及每一容量对应的标准误差的附表。这些表通常基于研究的目的是得到比例的估计值,以及比例的最大假设值($p=0.5$)是合适的这两个假定。不仅如此,这些表中列出的值,通常都未做过有限总体修正。这些表只能应用于非常有限的场合,在使用时必须备加小心!

分析研究的高效样本量的计算问题远远超过本抽样教科书介绍的范围。对这一问题感兴趣的读者,不妨首先读一读李普希(Lipsey, 1989)有关功率分析(power analysis)的著作。有一个研究实例可能对功率分析这一概念作了比较确切的解释。

有人曾经做过一项研究,旨在探测白人和少数民族在犯有同样的罪行时,刑期有什么不同。该项研究用下面的公式来分析白人和黑人的样本均值差:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{(s_{\bar{x}}^{\ 2} + s_{\bar{x}2}^{\ \ 2})^{1/2}}$$

在这一研究中,研究者决定如果差别是不存在的,那么他们犯错误的风险将是5%,即在 100 次试验中,可能有 5 次发现差别存在(即 $t=1.96$)。

不仅如此,我们可以从以往的研究中假设,黑人犯强奸罪的刑期的标准差是32.3 个月,白人则是 29.3 个月。最后,我们假设样本是等容量的($n_1=n_2$)。在采用了这些假设之后,如果探测的刑期差高达 16 个月,那么每个样本的个案量只需达到 30 个就可以了。然而,假如预期的刑期差只有 3 个月的话,那么为了能观察到这样大小的差别,每一个样本各需要 775 个观察。对于探测不同大小的期望的差别所需的样本量已在下面的表 7.2 中列出。

表 7.2 探测刑期差所需的样本量

每组样本量 （总样本量 = $n\times2$） n	刑期差 （月） $\overline{X}_B-\overline{X}_W$
30	16
50	12
75	10
115	8
295	5
775	3

表中的n 是每组的样本量，
$\overline{X}_B$是黑人的刑期的均值，
$\overline{X}_W$是白人的刑期的均值。

这些估计值清楚地显示，期望的效应越小，探测效应所需的样本就越大。这一关系可推广用于其他的分析研究。有时计算效率的特定公式可能会变得非常复杂。李普希（Lipsey ,1989）为我们提供了一个非常好的参考框架，引导研究者借助那些对设计非常敏感的因素去探测与样本容量有关的种种关系。在效率问题的计算比较错综复杂时，我们不妨去寻求抽样专家的帮助。

样本容量设计的实质

虽然对高效样本量具有最为直接影响的是样本量，但是研究设计对高效样本量的影响同样也是不可忽视的。高效样本量的计算都基于简单随机样本这一假设。如果样本设计的抽样方法并非简单随机抽样，那么高效样本的容量则也可能与简单随机抽样的高效样本量有所不同。在使用整群抽样时，抽样变异会有所增加；而在使用分层抽样方法时，抽样变异则会有所下降。

设计效应（deff）是一种直接表达设计对抽样变异性的影响的方法。在计算高效样本的容量时，我们可以将设计效应乘以期望的抽

样方差($s_{\bar{x}}^2$),以便对之进行修正。设计效应是设计的抽样方差与样本方差的比率(Kish,1965;Sudman,1976)。在实际工作中,为了使之能与标准误差进行比较,我们更多地使用的是设计效应的平方根。

在进行样本设计之前,我们必须先把设计效应代入高效样本容量的计算式,而要做到这一点,我们需要得到有关期望的设计效应的信息。不言而喻,如果有研究者能为我们提供这样的估计值,那是最好不过了。至于分层样本,均值的设计效应可能在 0.5 到 0.95 之间。而实际的效应将取决于分层的数目、分层变量和研究变量之间的相关程度。

不难想象,整群样本的设计效应一般都大于 1。在一般情况下,它在 1.5 到 3.0 之间。这样一个大小的范围,显然会对高效样本的大小有相当的影响。效应的估计值的确定取决于特定的设计本身所具有的一些性质。群的数目、群内成员的同质程度和分层的使用与否都会对实际的设计效应有很大影响。

多级样本的设计效应也应该做修正:"多级随机样本的抽样误差几乎都大于没有任何约束的随机抽样,在第一级(也可能在随后的各级)进行分层固然可以对因分级而引起的额外的抽样误差的减少有一定作用,但它永远也不可能将它完全去除。"(Stuart,1963,p. 89)斯图加特给我们提供了一条抽样误差增加 1.25 到 1.50 之间的经验法则。我们应该将这些数字的平方乘以样本方差。

此后出版的有关著作(Kish & Frankel,1970;Frankel,1971)给出了各种各样的多级样本的设计效应的平方根和各种各样的估计值。我们在这里特别要给读者介绍凯思和弗兰克尔的有关结论:"机器计算的标准误差,是基于简单随机样本这一假定的,(对于多变量分析)它显然是被低估了的";且"已经证明,设计效应是可以估计的和能反映回归系数的标准误差的量的"(Kish & Frankel,1970,p. 1073)。

在经验调查中,我们发现均值的设计效应大于回归系数。斯图加特(Stuart ,1963)给出了一个大致的估计,可以证明它所给出的均值设计效应的平方根还是比较精确的。而回归系数的设计效应的平方根似乎在 1.06 到1.30之间。在高效样本量计算公式中使用设计效应可能意味着使算出的样本量有很大的增加。

子总体分析

迄今为止,我们对样本容量的考虑始终基于所做的分析涉及整个目标总体这样一个假定。而在许多情况下,我们在对整个总体感兴趣的同时,也可能对它的某些子总体同样感兴趣。例如,在高龄老人研究中,可能有某一研究者希望能对女性老人单独进行研究,而另一位研究者则希望对州内的不同的地区一一进行研究。通常,子总体分析的精度总是低于将整个样本作为一个群体进行研究的精度。子总体所含的个案比较少,因而子总体的分析的抽样变异便会有所增加,尽管抽样变异的增加会因子总体的标准差比较小而有所抵消。

我们用一个例子来阐述子总体分析可能产生的问题。在这一例子中,项目分析人员准备估计社会服务机构所属的有关部门的开放时间,时间的量度单位是星期(表 7.3)。高效的总样本量为 1 620 个左右。进行一个八个区的、容量相等的子总体分析,将会产生一个含 203 个单位的子样本。区的子样本的标准误差是 2.64,而与之对应的总样本的标准误差只有 0.9,区子样本的标准差几乎是总体的 2.9 倍。区的 95% 的置信区间的总的大小是 10.3(2.64 × 1.96 × 2)。如果得到区的估计值的确很重要,那么研究者就必须考虑,诸如这样的 ±5 区间,就研究目的而言是否已经足够精确。如果不够,研究者必须考虑增加样本量。但与此同时研究者也必须考虑将精度提高到可以接受的水平所需的总样本量的费用问题。

表 7.3 样本容量和子总体分析(分析 8 个等容量的区)

$n = 1\,620$　　$s = 37.6$　　$s_x = 0.9$

$n_d = 1\,620/8 = 203$

$s_{d\bar{x}} = 37.6/(203)^{1/2} = 2.64$

其中,n 是总样本量,

n_d是每个区的样本量,

s 是标准差的估计值,

s_x是整个样本的均值的标准误差,

$s_{d\bar{x}}$是区的子样本的标准误差。

不合格和无回答的修正

在选择样本容量的时候,研究者必须牢记,样本精度是根据目标总体中能实际收集到数据的成员数估计的。我们有时无法从样本中的某些成员那里收集到有用的信息,其原因不外乎以下两种:

- 样本框中包含不合格的成员,
- 无回答。

不合格是指那些虽然列在了样本框内,但并不属于目标总体的成员。不合格成员使实际的样本量减少,从而导致抽样变异性的上升。例如在北卡罗莱纳州居民调查中,不难想象一个在北卡罗莱纳州工作和缴税的弗吉尼亚的居民会出现在该州的纳税人卷宗中。而这一个体对于该州居民的抽样框而言,便是不合格的。与此类似,在用随机数码拨号法对一个特殊群体进行调查时,便会有许多电话打到了宅内没有目标总体成员的宅邸,这样的号码便会因此而从样本筛选掉。佛罗里达调查便是一个例子。

导致无回答的原因有很多,包括回答人无法联系到和拒绝回答等。无回答会造成样本的非抽样偏倚,因为它会造成总体中的一部分成员在样本中的代表性过低。在下一章,我们将就无回答偏倚的评估问题展开讨论。在这里我们将给诸位提供一种对高效样本量进行修正的方法,以对无回答给抽样变异性带来的问题进行弥补。

我们可以将高效样本量除以合格成员的比例乘以回答人比例之积来对不合格和无回答造成的影响进行弥补。在样本框中合格成员的比例为0.95,预期样本中回答我们的问题的回答人的比例为0.85时,若高效样本量为1 620,那么对之进行修正之后,其初始样本量约为2 006。

$$
\begin{aligned}
n' &= \frac{n}{e \times r} \\
&= \frac{1\ 620}{0.95 \times 0.85} \\
&= 2\ 006
\end{aligned}
$$

式中，n'是修正后的样本量，

n 是高效样本量，

e 是清单中合格成员的比例，

r 是期望的回答问题的回答人比例。

回答率越高，作为高效样本量修正结果的初始联系数（初始样本量）就越少。为了提高回答率所做的深入的追踪的费用都比较昂贵，而较小的初始样本量则可部分地抵消由此而产生的额外的费用。

费　用

现在我们已经可以对数据收集所需的费用做一个大致的估计。我们应该从以下几个方面来考虑所需的各种费用：

- 得到一个抽样框所需的费用，例如为掌握面积概率抽样所必须的住房的实际地点而进行的实地调查的费用；
- 得到和使用样本选择所需信息（分层变量）所需的费用；
- 数据收集，包括整个样本的初始联系和访谈收集数据（交通费、电话费和人工费等）或管理使用调查工具（邮资、交通费和人工费等）的费用；
- 进行跟踪调查所需的费用；
- 对于已经完成的工具整理和编码的费用；
- 计算机分析所需的费用。

除了前两项费用之外，其余费用都随所选的样本量或得到的答案的数目的变化而变化。在计算费用的时候，跟踪调查所需费用尤其重要。在跟踪调查方面进行的投资可以提高回答率，进而降低选择的样本量。它还可以降低与初始联系数有关的费用，并通过降低潜在的非抽样偏倚而免除对无回答偏倚评估所需的费用。例如，在上面引用的例子中，回答率若能从85%上升到95%，那么修正的样本量便可以从2 006 降到1 795。那些试图与额外的211 个样本单位联系有关的各种步骤，如得到他们的具体地址或电话号码、邮寄问卷、打电话约时间等的费用，可能比深入跟踪更贵。不仅如此，深入

跟踪还可以降低非抽样偏倚。

信　度

高效的样本量并非总是可信的样本量。信息的用户常常会对样本的信息有所疑惑,因为在他们看来,这些信息所依据的个案太少。有时,听取信息的那些人自己对地区之间存在的差异或地方的独特性有一定了解。在遇有这样的情况是,我们在设计时可能需要考虑配置一个比较大一点的样本。

有时,上面提到的那些问题是作为总体大小的函数出现的。而上面介绍的那些高效样本量的计算方法基本上都未考虑总体的大小。实际上,总体的大小只是在纠正因子中才使用。然而,诸如样本容量应该是总体容量的一个百分比——通常是10%——这样的观点的确存在。这样的观点并不正确。在一般人口总体的民意调查或选民调查中,通常使用的样本量在1 500 到2 500之间。尽管有些时候人们会提出诸如1 500 个个人怎么能代表整个国家的公民这样的疑问。媒体的使用和民意测验结果的精确性已经使很多诸如这样的怀疑冰释,不过也许有矫枉过正之嫌。

有时用于媒体研究和其他特殊人口总体研究的样本容量较小会被人认为不可信。对样本信度的怀疑会因为在随机选择过程无法做到样本单位的按比例分配而加剧。某些立法区比例的缺乏可能被看做立法和政策制定过程中样本信息的缺失。我们将对过分强调样本比例要反映总体比例的错误观点进行讨论。

例如,在一个评估项目中,我们抽取了一个登记为成年人居住的60 户的住户样本进行考察和数据收集,但项目管理者却拒绝接受评估的结果。而对全体400 个住户重新普查之后,我们发现有问题的家庭在观察的几个变量上存在的差异不超过3%。

在有政策影响的研究中,对于样本信度的责难是不可能消除的。事先进行周密的计划和对那些可能影响样本信度的因素予以足够的关注也许可以抵挡那些不正当的责难。研究者应该倾听那些对调查结果有所期盼的人们的意见,并在设计阶段便使它们能在样本中有所反映。人们反映的问题,如分地区确定样本量也许可以使研究者

通过更改设计方案来排解那些可能出现的责难。

小总体抽样

研究者有时会需要从一些比较小的总体进行抽样。在这样的场合,样本究竟应该多大？诸如县、监护官或拉丁语课程的注册学生这样的总体一般都不太大,但对有限的资源而言,我们仍然无法对诸如这样的整个总体收集数据。然而,可能的样本量对得到可信的结果似乎又过小。在遇有这样的情况时,常常需要考虑对目标总体的成员采用确定性选择法(certainty selections),以使样本具有代表性,因而有可信性。由此得到的结果必须做适当的加权,以对选择的概率做出说明。

另一种可以用于小样本的方法涉及数据中的异常值(outlier)分析问题。小样本特别容易出现异常值。异常值可以对样本估计值产生异常的影响。这一问题的困难之处在于什么时候一个观察的值才是异常的和什么时候它才是总体的合理的代表？用于异常值的统计分析,并能提供可信和稳定的统计值的方法近年来已经有了很大的发展(Andrews, Bickel, Hampel, Huber, Rogers, & Tukey, 1972; Barnea & Lewis, 1984)。

其中折刀法(jackknifing)不仅非常有用,而且简单易行(Efron, 1982)。折刀法反复地迭代移除单个观察,并计算样本估计值,直至每一观察都从一次计算中被移除过为止。然后我们再对得到的一系列估计值排序并绘制散点图,显示估计值对任一单个观察的敏感度。图 7.2 便显示了一个样本均值的敏感度。敏感度曲线展示了一个 128 个单位的区域,显示了极端值对数据的强烈影响。

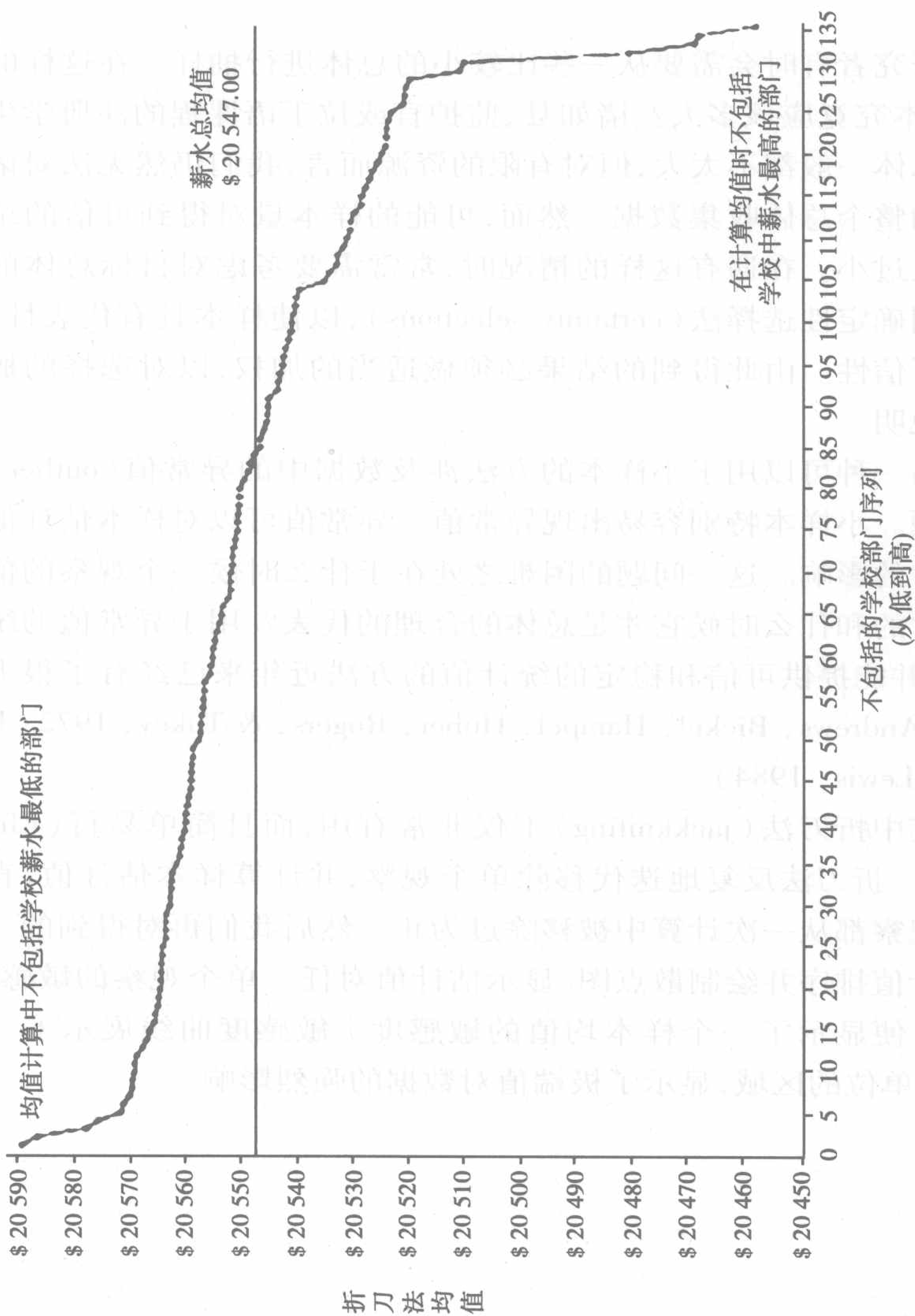


图7.2 使用折刀法的敏感性曲线

小 结

样本量的选择取决于可以容忍的抽样变异性的大小、重要的变量的变异性、设计效应、子总体分析的需求、不合格成员和无回答的数量、费用和信度等因素。在对各种可供我们选择的样本设计方案进行评估时,我们应该对这些因素予以综合考虑。本章所提供的公式使研究者得以在考虑这些因素之后,估计调查所需的样本量。

第8章 抽样后选择

Postsampling Choices

完成数据收集并不意味着所有与抽样有关的问题都迎刃而解，仍然有一些与之有关的问题需要我们进行处理。这些问题关系样本的设计和设计方案执行过程中出现的问题，主要有：

- 权的使用；
- 无回答评估；
- 陈述数据(presentation of the data)。

下面我们对这些问题逐一进行讨论。

权的使用

在样本并非等概选择的时候，通常都需要进行加权。不等概选择可能会产生样本偏倚。正因为如此，我们有必要回顾抽样设计中我们所作的每一个选择，以确定在什么地方偏离了等概性。可能产生不等概问题的设计环节是各种各样的。我们将前面几章中提到的与之有关的问题概括为以下几个方面：

- 抽样框含有重复的条目；
- 登录的条目是群，例如抽样框登录的是户；
- 采用了不成比例抽样方法。

我们应该分别就不同的原因来考虑加权问题。在某些时候，权的使用取决于某一特定的问题的分析单位。

权和分析单位。例如,在北卡罗莱纳州居民调查中,权是用来对从户群中选择的个人进行补救。然而,在问题本身关系户而非个人的时候,我们就无需再加权。无独有偶,在出院的精神病人调查中,在研究所关注的问题是涉及精神病患者个人的一些特质(如进入某一机构的次数)时,由多次出院引起的抽样框的重复便会造成不等概。有关出院过程的问题是基于出院的等概的选择。例如,一个有关出院比例的问题涉及的是出院前有关社区招募照料出院病人的社会工作者的计划问题,这一问题便与加权无涉。

分层后加权。为了对样本与目标总体中的特质的分布之间的偏差做出修正,我们常常需要考虑进行加权。这种类型的加权叫做分层后加权。表 4.3 显示了北卡罗莱纳人口的样本和普查的估计值之间的差异。二者在男女比例上存在的差异,已大到使我们不得不考虑男性的代表性是否不够这一问题。如果研究者认为情况确实如此,那么我们需要对样本观察值进行分层后加权:

$$w = \frac{p_p}{p_s}$$

式中,p_p是总体的比例,

p_s则是样本的比例。

该公式可以用于单变量的修正,或基于目标总体双向或多向列联表的单元格比例的修正,如果这些比例是可以得到的话。

在分层后加权,或任何解决那些问题的加权应用之后,我们应该对加权对结果产生的影响进行分析。如果权的影响是微不足道的,那么我们就有理由把权删除,使分析简单化。

在采用分层后加权之前,还有两个问题应当进一步引起我们的注意。首先,我们应当认真斟酌总体数据的精确性。在佛罗里达高龄老人研究中,我们最初得到的总体的估计值来自四年以前的普查的估计值。85 岁以上老人的百分比,在普查估计值和样本估计值之间最初观察到的差异在 4 个百分点以上。老年人移民到佛罗里达和人口老龄化也许可以在一定程度上对这一差异做出解释。我们又进一步考察了来自 1985 年的普查数据(表 4.5),发现这一差异降到了

2.7%，已经在我们期望的范围之内。因而，我们认为没有必要再用加权来弥补这一差异。

其次，后分层法虽然不是解决无回答问题的万应灵丹，但它却可以用于样本和目标总体特征值之间出现的在我们意料之中的随机差异的修正。在用后分层法做无回答修正时，我们假设无回答的回答人将以与自己有着类似人口学特征的回答人相同的方式回答问题。这一假设必须在经验上予以证明。无回答者至少有一点与回答人是不同的，那就是他们选择了不回答。调查的问题对回答人的重要性和他们在调查期间是否有时间等，都可能是造成回答问题与否的重要原因。无回答问题的评估是下一节所要讨论的问题。

无回答评估

无回答可能会造成非抽样偏倚，在数据收集工作完成之后，我们不可对这一问题掉以轻心。无回答问题的影响可能会是很严重的。无回答的影响与抽样框不完整所造成的影响颇为相似，它会使一部分目标总体从样本中丢失。样本无法代表那些丢失的个体，因而不能成为总体的一个精确的模型。

为了评估无回答者对样本估计值的影响，卡尔登建议将他们看做一个层（Kalton，1983）。这时，计算分层样本的均值的公式是：

$$\bar{x} = \left(\frac{n_r}{n}\bar{x}_r\right) + \left(\frac{n_n}{n}\bar{x}_n\right)$$

式中，n_r是回答人的数目，

n_n是无回答者的数目。

从上式可知，整个样本的均值是回答人的均值和无回答者的均值的加权平均数。

我们可以用这一公式设计一种处理无回答问题的策略。首先，最好的策略发端于编制一个最大限度地减少无回答的计划。无回答的比例越小，它对整个平均数的影响就越小。减少无回答的计划，取决于目标总体、数据收集的方法和可用于降低无回答的各种工作的资金的多寡这三者。

与抽样设计的大多数问题一样，最大限度地减少无回答的计划的制订，应该在设计方案最终确定之前。为了减少无回答，也许采用比较小的样本并将资金用于各种能得到更多的个案的数据收集工作的做法不失为明智之举。在对精神病患者出院的研究中的研究者，使无回答降到了 0.86%（3/350）。他们的做法是周密的计划和对最初的无回答者进行深入的跟踪调查。尽管深入的跟踪调查对于一般人口总体的样本也许不太现实，但是我们还是不难发现它对于降低无回答起到的作用。我们不难在许多调查教科书中，包括在福勒（Fowler，1984）、拉夫拉卡斯（Lavrakas，1986）、迪尔曼（Dillman，1978）及苏德曼和布拉德伯恩（Sudman & Bradburn，1982）等人的著作中发现各种深入跟踪调查的策略和方法。一般讲，跟踪联系的策略和方法越个人化，得到无回答者的回答的可能性就越高。

在权衡初始样本量和跟踪调查的深度，并作出抉择之后，仍然存在相当大的产生偏倚的可能性。10% ~ 20% 的无回答可能会产生显著的偏倚。一种人们比较喜欢使用的用于评估无回答影响的方法是，设法得到无回答者样本的数据。这种方法涉及如何从无回答者层进行抽样，进而用面对面访谈或电话调查收集样本数据。

这时，数据收集的数量与最初使用的工具相比，可以显著减少。为了限制数据收集的数量，我们应该选择那些最感兴趣的和那些有理由怀疑在无回答者和回答人之间存在明显差异的变量。回答人层的分析也许可以给我们提供有用的信息，使我们得以了解哪些自变量很重要，哪些预测变量的分布也许会对整个估计值产生影响等。为了估计无回答者对整个估计值的影响，我们可将样本结果用于上面的公式。抽样误差应该用分层样本的公式进行计算。

对无回答者样本进行的调查研究应该是对跟踪调查策略的补充而不是取代。在处理无回答问题时，我们首先应该考虑的问题是如何最大限度地减少无回答。不采取任何减少无回答的措施，任其发生，必定会因为要涵盖一个有偏的估计值而加大置信区间，从而使总误差激增。在前一章介绍的无回答修正的样本量只是总误差的一个方面，即对抽样偏倚进行了纠正。在因为时间和资源的限制而无法进行无回答样本的调查研究时，有两种使研究者得以了解无回答的影响的方法供他们选择。第一种是对第一批得到的答案与第二和最后一批得到的答案的模式进行分析，看看几批答案之间是否存在差

异。如果不同"批"的答案之间不存在差异,则说明无回答偏倚可能比较小。这一分析基于的假设是:后面的回答者可能与无回答者有着共同的特征。如果确有差异出现,那么最后一批答案便可用于计算加权均值的公式,以对无回答组进行估计。

最后一种评估无回答的影响的方法是用分层加权公式平均确定无回答者可能展现的推翻研究结论的模式。这个公式可重写为下面的形式。

$$\bar{x}_n = \left(\bar{x}_c - \frac{n_r}{n}\bar{x}_r\right)\frac{n}{n_n}$$

式中,$\bar{x}_n$ 是无回答层推翻结论所必需的均值,

$\bar{x}_c$ 则是可能会推翻结论的总均值。

例如,在一次测验中,回答人的平均得分是百分制的 94 分,那么为了证明成绩有所提高,将测验成绩的低限(总平均)设为 88 分就足够了。如果我们得到回答率是 90%,那么为了推翻成绩有所提高的结论,无回答者的平均分必须低于 34 分。如果回答率降到 80%,那么无回答者的平均分即使达到 64 分,也会改变原有的结论。而 75% 的回答率可能需要无回答者的平均得分在 70 分以下才能改变原有的结论,而这一分数肯定位于它所基于的分布的合理的范围之内。

综上所述,我们不难理解,无回答远不只是一个回答人在回答了其他有关题项,却拒绝回答某一题项的问题。这个问题与之有很大差别。从已经得到的信息中推断某些有关的值的各种方法统称为配值(imputation)。可供我们使用的配值法有很多,包括"热点甲板"(hot deck)法和回归分析法(Kalton & Kasprzyk,1982)。

陈述数据

本节不打算在技术报告的撰写问题上多费笔墨,而准备用较多的篇幅介绍如何比较精确地陈述样本信息。近年来,样本信息概括报告在使用户和公众更多了解样本的质量方面已经起了很大作用。在报纸和电视广播发表的民意测验结果报告中已经常常出现诸如"误差范围"这样的字眼。虽然,只有在某些场合的报告中才会对标

准差做陈述，但置信区间一侧的宽度（$\pm t(s_{\bar{x}})$）似乎已在这些报告中广泛使用。公众已经开始理解“抽样变异”这一概念的含义，或在许多情况下，公众在消化来自民意测验的信息时，至少已经有了一些信息可供他们对它加以考虑。不过，迄今为止，对误差的其他组成部分的讨论仍付诸阙如。

在这一节我们将向读者比较介绍许多数据陈述报告的方法。我们常常会看到陈述报告在未对总误差做任何讨论的情况下陈述报告点估计值。有些报告将标准差列在了表格中，或在显著性检验中对之作了明确的陈述。我们也常常会看到，在数据陈述报告中没有对潜在的偏倚进行任何讨论。

任何信息的使用都会有一定的风险，那就是我们所使用的信息有可能被证明是不够精确的。在理论检验时，科学家必须确定数据对理论进行的检验是否足以消除人们在这一领域内提出的其他种种质疑。在研究报告中陈述的种种有关抽样的选择应该清晰到能令其他训练有素的研究者确定这些信息是否有效和可信的程度。

政策的制定者在使用有关信息时同样也要冒一定的风险。项目官员、行政管理人员和选出的公务人员也许并不能充分理解这种风险。对于这样一些读者和听众，陈述报告必定不能只是抽样设计时做出的种种选择和样本实施过程的简单复述。陈述报告不应该是技术性的附录，而应该是实施过程的总结和概括。这一报告需要陈述的是客观事实和主观判断。

在这份报告中，必须对总误差的每一组成部分进行讨论。但这种讨论不要从理论角度展开，而要从它们对结果的影响展开。因为并非所有的影响都是可以量化的，因此研究者本人的主观判断是十分必要的。目标总体和研究总体之间存在的差异、无回答和标准误差都应该展开讨论。对来自抽样偏倚或源自其他方面的细小的偏倚应该留到技术性更强的讨论时讨论。

在确定什么样的潜在误差应该报告和怎么样报告时，研究者应该设身处地从读者和听众的角度进行考虑。如果潜在误差有可能改变读者和听众将要采取的行动的结果，那么研究者就有责任提供这方面的信息，以便他们能对行动的风险做出判断。然而困难之处常常在于，我们难以确定，对那些外行的读者和听众，我们的信息究竟应该提供到什么程度？而更难确定的问题是，究竟什么样的信息才

能使那些外行的读者和听众明白？含有置信区间具体范围的统计图和围绕这些范围来讨论统计值，可能对抽样变异性理解很有用处。在本章前面部分提到的，基于分析结果陈述的偏倚可能造成的影响，应该在讨论研究发现的时候把它作为一种预示加以阐述。

在最后的分析中，研究者给读者和听众提供的信息的多寡和精确程度将决定他们对所冒的风险承担的责任的大小。在得到的信息已足够对使用研究结果所冒的风险做出判断时，信息使用者就应承担使用这些信息的责任。在研究结果的不确定性并未得到应有的揭示时，研究者就应分担部分责任。

结 论

重要的问题在于我们必须把实用抽样设计的基本思想贯穿于整个研究设计、具体实施、数据分析和调查报告撰写的所有阶段。在研究的某一阶段做出的选择将对以后的选择和程序有所影响。明白无遗地将所有抽样的选择组合成一个完整的研究过程，将有助于我们把注意力投向总误差。在资源有限的前提下，降低总误差是实用抽样设计要实现的最终目标。

总误差由三个部分组成：非抽样偏倚、抽样偏倚和抽样变异。每一部分都会对研究结果的精确性有影响。诸如抽样框的选择、样本容量的选择这样的设计选择将会直接对总误差的大小产生影响。而那些评估性的程序，如无回答样本调查或标准差计算，则会为我们提供有关我们从样本中发现的结果中实际存在的误差的信息。研究者应当帮助调查结果的用户了解提供给他们的研究发现的精确程度，以使他们能对使用这些结果可能要冒的风险做出判断。

参考文献

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances*. Princeton, NJ: Princeton University Press.

Baker, T. L. (1988). *Doing social research*. New York: McGraw-Hill.

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data*. (2nd ed.). New York: John Wiley.

Bradburn, N. A., & Sudman, S. (1980). *Improving interview methods and questionnaire design*. San Francisco: Jossey-Bass.

Burnam, M. A., & Koegel, P. (1988). Methodology for obtaining a representative sample of homeless persons: The Los Angeles skid row study. *Evaluation Review*, 12, 117-152.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental design for research*. Chicago: Rand-McNally.

Cook, D. C., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Czaja, R., Blair, J., & Sebestik, J. P. (1982). Respondent selection in a telephone survey: A comparison of three techniques. *Journal of Marketing Research*, 21, 381-385.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: John Wiley.

Dillman, D. A., & Tarnai, J. (1988). Administrative issues in mixed mode surveys. In P. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, & W. L. Nicholls (Eds.), *Telephone survey methodology* (pp. 509-528). New York: John Wiley.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling*

plans. Philadelphia: Society for Industrial and Applied Mathematics.

Fowler, F. J., Jr. (1984). *Survey research methods*. Beverly Hills, CA: Sage.

Frankel, M. R. (1971). *Inference from survey samples: An empirical investigation*. University of Michigan, Institute for Social Research, Ann Arbor.

Grizzle, G. A. (1977). *North Carolina Citizen Survey, 2: How the survey was conducted and what it cost*. Raleigh: Office of State Budget and Management.

Hess, I. (1985). *Sampling for social research surveys 1947-1980*. Ann Arbor: University of Michigan.

Holt, M. M. (1977). *SURREGR. Standard errors of regression coefficients from sample survey data*. Research Triangle Park, NC: Research Triangle Institute.

Joint Legislative Audit and Review Commission. (1979), *Deinstitutionalization and community services*. Richmond: Virginia General Assembly.

Joint Legislative Audit and Review Commission. (1986). *Deinstitutionalization and community services*. Richmond: Virginia General Assembly.

Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills, CA: Sage.

Kalton, G. Unpublished manuscript, 1986. *Models in the practice of survey sampling*. University of Michigan, Institute for Social Research, Ann Arbor.

Kalton, G., & Kasprzyk, D. (1982). Computing for missing survey responses. *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 22-31.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Kish, L., & Frankel, M. R. (1970). Balanced repeated replications for standard errors. *Journal of the American Statistical Association*, 65, 1071-1094.

Kraemer, C. H., & Thiemann, S. (1987). *How many subjects? Statis-*

tical power analysis in research. Newbury Park, CA: Sage.

Lavrakas, P. (1986). *Telephone surveys*. Newbury park, CA: Sage.

Lipsey, M. W. (1989). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Mark, H., & Workman, J., Jr. (1987). Populations and samples: The meaning of "statistics." *Spectroscopy*, 2 47-49.

McKean, K. (1987, January). The orderly pursuit of pure disorder. *Discover*, 72-81.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: John Wiley.

Office of State Budget and Management. (1982). *North Carolina citizen survey: Highlights* 1981. Raleigh: Office of State Budget and Management.

Office of State Budget and Management. (1983). *North Carolina citizen survey: Highlights* 1982. Raleigh, NC: Office of State Budget and Management.

O'Rourke, D., & Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, 20, 428-432.

Raj, D. (1972). *The design of sample surveys*. New York: McGraw-Hill.

Rog, D. J., & Henry, G. T. (1986). *A community profile of the deinstitutionalized*. Unpublished manuscript, American Psychological Association.

Rossi, P. H., Wright, S. D., Fisher, G. A., & Willis, G. (1987). The urban homeless: Estimating composition and size. *Science*, 235, 1336-1341.

Shah, B. V. (1981). *SESUDAAN: Standard errors programs fcr computing of standardized rates from sample survey data*. Research Triangle Park, NC: Research Triangle Institute.

Skidmore, F. (1983). *Overview of the Seattle-Denver income maintenance experiment: Final report*. Washington, DC: Government Printing Office.

Smith, T. M. F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society*, 139, 183-204.

Stuart, A. (1963). Standard errors for percentages. *Applied Statistics*, 12, 87-101.

Stuart, A. (1984). *The ideas of sampling*. New York: Oxford University Press.

Sudman, S. (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61, 749-771.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions*. San Francisco: Jossey-Bass.

Stutzman, M. (1985). *Florida's 75 + population: A baseline data sourcebook*. Tallahassee: Florida State University.

Troldahl, V. C., & Carter, R. E. (1984). Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, 1, 71-76.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Williams, J, A., Jr. (1982a). *North Carolina citizen survey: Overview, Fall* 1981. Raleigh, NC: Office of State Budget and Management.

Williams, J. A., Jr. (1982b). *North Carolina citizen survey: Technical report, Fall* 1982. Raleigh, NC: Office of State Budget and Management.